The Value of Data

Simone Galperti
UC, San Diego

Aleksandr Levkun
UC, San Diego

Jacopo Perego
Columbia University

June 11, 2021

Preliminary and Incomplete

Abstract

Data is an essential input of many modern industries. Yet, its value is hard to establish, since formal markets for data are still lacking. We study what value different types of data have for a principal—e.g., an e-commerce intermediary—who uses it to mediate the interaction between multiple agents—e.g., buyers and sellers. Our solution formulates this mediation as an information-design problem and uses linear-programming duality to characterize the principal's willingness to pay for each type of data. This reflects externalities between datapoints, which arise from how the principal optimally garbles them. Building on this, we study how the principal values information that refines existing datapoints, which we show can be zero or even negative. Our work establishes basic properties of the "demand for data," a necessary step towards a full analysis of data markets and their welfare properties.

JEL Classification Numbers: C72, D82, D83

Keywords: Data, Information, Duality

We are thankful to S. Nageeb Ali, Alessandro Bonatti, Laura Doval, Elliot Lipnowski, Joel Sobel, Denis Shishkin, and seminar participants at Cambridge University, Columbia University, Penn State University, Pittsburgh University & CMU, UC Davis, and UC San Diego for useful comments. All remaining errors are ours.

The first step toward valuing individual contributions to the data economy is measuring these (marginal) contributions. (Posner and Weyl, 2018, p. 244)

1 Introduction

Data is the "new oil" in modern economies and a topic of major debates about privacy. Yet, data is often not traded in formal markets, nor are individuals compensated when their personal data is used. Data is usually collected for free or, at best, bartered in exchange for online services. Either case may result in significant inefficiencies and misallocation. Many scholars and policymakers view establishing functioning data markets as essential for the digital economy to bring prosperity and stability to societies (Lanier, 2013; Posner and Weyl, 2018; Arrieta-Ibarra et al., 2018). A key challenge is to determine the value of an individual's specific data (Acquisti et al., 2016; Posner and Weyl, 2018). Is one consumer's data more valuable than another's for an e-commerce platform? If so, how much should each be paid?[1]

This paper breaks new ground in this direction in two ways. First, we study the value of data in settings where a third party uses it to mediate interactions between agents with conflicting interests. Besides e-commerce platforms, examples include matching markets (like ride-sharing and navigation services) and auction-based markets (like ad auctions). Second, we characterize the value of data given its *specific* realization. This is a non-trivial task that requires a new approach. In short, data grants the mediator an informational advantage, which she exploits in complex ways that create externalities between the data of different interactions.[2] Our solution involves modeling this mediation as an information-design problem and leveraging its structure as a linear program. Our values represent the mediator's willingness to pay for data and a benchmark for fairly compensating its sources. Building on this, we also study the value of information in mediation problems. Our results shed new light on the demand side of data markets.

Consider an example. An online platform mediates trade between a monopolist and a population of buyers. For each buyer, it owns data consisting of a record of personal characteristics, including a unique identifier. This record reveals the buyer's valuation of the seller's good, which

[1]These issues have also been debated in policy circles. The World Economic Forum (2019) raises similar questions, "How do we determine if one person's data packet is more valuable than another's, even if they both relate to the same set of online behavior? How do we value quantity vs quality of data?"

[2]These externalities differ from those highlighted by Acemoglu et al. (2021) and Bergemann et al. (2021) as explained below.

is produced at zero marginal cost. We refer to each buyer and his or her record as a *datapoint* in the platform's database. This contains two types of datapoints, labeled by $\omega \in \{1, 2\}$ and corresponding to whether the buyer's revealed valuation is 1 or 2. [3] In total, 3 million datapoints have $\omega = 1$ and 6 million have $\omega = 2$. The seller knows only the database composition $q = (3M, 6M)$. The platform mediates the interaction between each buyer and the seller by disclosing a signal about ω so as to influence the price the seller charges (as in Bergemann et al., 2015). A classic question is how much value the platform derives from the *whole* database: This is the overall payoff from its optimal use. By contrast, we ask how much value it derives *individually* from each ω-datapoint, which we denote by $v^*(\omega)$. Note that each buyer-seller interaction is the basic unit of the platform's problem and the vehicle through which it derives value from its data. Finding this value would be standard if the platform itself were the seller and thus solved the decision problem of maximizing profits with each buyer: In this case, $v^*(\omega) = \omega$.

To highlight the role of conflicting interests in mediation problems, suppose the platform maximizes the consumer surplus of each buyer, which goes against the seller's profits. To do so, it sends signal s' for all 1-datapoints; for each 2-datapoint it sends s' and another signal s'' with equal probability. After s', the seller is indifferent between a price of 1 or 2, but picks 1 in favor of the platform; after s'', the optimal price is 2. The platform's payoff equals 0.5 units of surplus on average for each 2-datapoint and zero surplus for 1-datapoints. Are these payoffs the actual value the platform derives from each datapoint? The answer is no. Perhaps counterintuitively, here the most valuable datapoints are those yielding the lowest payoff for the platform. Indeed, note that 2-datapoints yield a positive surplus only when combined with 1-datapoints in signal s'. The latter datapoints "help" to persuade the seller to charge a low price to some high-valuation buyer. Hence, they should not be worthless, even though each interaction with $\omega = 1$ by itself yields zero surplus. We will show that $v^*(1) = 1$, reflecting the fact that 1-datapoints exert a positive information externality on interactions with $\omega = 2$. By contrast, $v^*(2) = 0$ because 2-datapoints have to "repay" this externality to 1-datapoints.

This example illustrates three takeaways that we will develop in the paper. First, we cannot use the payoff the platform *directly* obtains from a datapoint to assess its value. This is in sharp contrast with decision problems, like maximizing profits. The reason is that, due to conflicting interests, the platform garbles datapoints to produce partially informative signals, thereby using data of one buyer-seller interaction to mediate other interactions. Second, the gap between the

[3] We use the word "type" to refer to one of the possible realizations of all the data about a given buyer-seller interaction. The set of possible types of a datapoint is formally analogous to the set of possible states of the world. We prefer the former terminology because in our setting all datapoints (i.e., interactions) physically exist at the same time, while usually we intend different states of the world to be mutually exclusive.

value and direct payoff of a datapoint reflects information externalities with other datapoints in the database. We characterize such externalities in terms of how a datapoint helps (or not) the platform mediate other interactions, and how the platform exploits the seller's incentives across interactions with its signals. Third, the optimal use of datapoints—hence, v^*—depends on the database composition q, as this determines the informational advantage and garbling options for the platform. For instance, if we swap the quantities of 1- and 2-datapoints (i.e., $q' = (6M, 3M)$), we get $v^*(1) = 0$ and $v^*(2) = 1$. We analyze how v^* varies with q and discover, among other things, the following general principle: For any mediation problem, the scarcer a type of datapoint is, the higher its value.

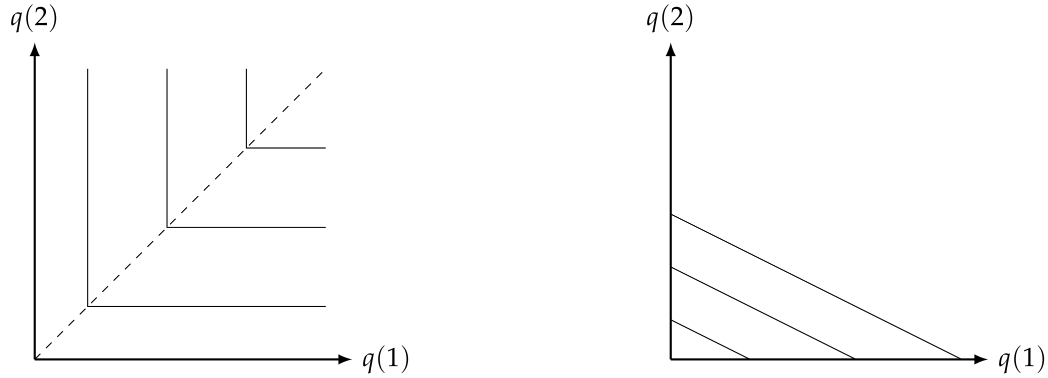
This example is an instance of the class of mediation problems we consider. A principal mediates interactions between multiple agents by providing them with information. She produces this information using data, which she already owns.[4] We view datapoints as physical inputs of this information production constrained by their quantity in the database. We use linear-programming duality to characterize the unit value v^* of these inputs, adapting classic work of Dorfman et al. (1987) and Gale (1989) to our class of problems. This approach allows us to handle the aforementioned complexities of assessing the value of data for mediation problems, highlighting key differences from settings where data is used to solve decision problems.

Another contribution of the paper is to show how v^* can guide mediators when acquiring more data. This can mean adding new datapoints to the database, or refining existing datapoints with more informative observations. For instance, our platform can expand its userbase and add datapoints (i.e., buyers and their characteristics) to its database $q = (3M, 6M)$. How much should it be willing to pay for new datapoints? The answer has to rely on v^* (not direct payoffs), so it is zero for 2-datapoints and one for 1-datapoints. If a new datapoint's type is uncertain, we simply take the expectation of v^*. More generally, v^* characterizes mediators' marginal rate of substitution between datapoints in the space of databases. Figure 1 illustrates our platform's indifference curves over databases. Interestingly, the same types of datapoints are perfect complements when the platform maximizes the buyer's surplus (panel (a)), but perfect substitutes when it maximizes the seller's profit (panel (b)).[5] In fact, we find that indifference curves are convex if and only if a mediation problem is non-trivial—i.e., its solution is *never* full disclosure, independently of q. This has immediate implications for which database a mediator would choose given a budget constraint and hence its demand functions for data.

With regard to refining datapoints, how much does the platform value learning more about

[4] In a related project, we analyze the case where the principal has to first elicit the data from its sources.

[5] It is easy to see that for profit maximization $v^*(1) = 1$ and $v^*(2) = 2$ independently of q.



(a) Maximize Buyer's Surplus.

(b) Maximize Seller's Profit.

Figure 1: Iso-payoff curves in the platform's example.

some buyers already in its database? To fix ideas, imagine some buyers have just joined the platform, which therefore knows their IDs but nothing about their valuation for the seller's good. Denoting this type of datapoints by ω°, we can apply our approach to find $v^*(\omega^\circ)$. We can then isolate the value of the *information* contained in an ω-datapoint for $\omega \in \{1, 2\}$ by netting out the value of the same interaction if the platform knew nothing about that buyer $\big($i.e., $v^*(\omega) - v^*(\omega^\circ)\big)$. The expectation of this difference gives the platform's willingness to pay for learning (WTPL) the buyer's valuation for ω°-datapoints. Importantly, these net values depend on q for mediation problems. In addition, WTPL can be *strictly* negative. A sufficient condition for it to be non-negative is that the platform learns independently across ω°-datapoints. For such refinements, we provide sharp conditions for WTPL to be zero, even though learning changes how the platform uses its data; moreover, WTPL decreases as more ω°-datapoints are refined. This may be reminiscent of standard decreasing marginal value of information. Here, however, the platform is not learning gradually more information about one specific datapoint; it is learning the same amount of information, but for an increasing share of ω°-datapoints. This process changes its informational advantage.

We derive these properties for any type of refined datapoints and any mediation problem. Key to this is characterizing the principal's preferences over a sufficiently rich space of databases, which we call *complete*. In short, given any database in this space if the principal can refine some datapoint by acquiring better observations, then the resulting database is also in the space. We can then view acquiring data—whether to add or refine datapoints—as moving inside this space and evaluate its effects using the principal's overall preference as pinned down by v^*.

More broadly, we hope this paper can speak to several issues in the digital world. Studying welfare effects of policies requires getting demand functions for data right. We show that it is not enough to focus on the data-user's direct payoff from each datapoint, such as revenues, sales, etc. The payoff consequences for other involved parties also matter, to measure the externalities between datapoints. We also offer a benchmark to assess how privacy protection affects the value of data: Our principal's direct access to it is akin to no protection at all. A better understanding of the value of people's data may help improve on the status quo of free data. Arrieta-Ibarra et al. (2018) list some of its pitfalls: jobs loss, wealth shift to the top, underproduction of the high-quality data that fuels AI productivity growth, and erosion of personal dignity. Lanier (2013) argues that compensating people for their data contributions to AI automation may be essential to save the middle class and ultimately democracy. Recently, "data unions" have emerged to intermediate individual sources of personal data and its user, which requires figuring out how much each should be paid.

1.1 Related Literature

Our work builds on the literature on information design (see, e.g., Bergemann and Morris, 2019, for a review). We formulate our mediator's "data-use" problem as an information-design problem and, as in Bergemann and Morris (2016), we use the revelation principle to express it as a linear program. Our analysis of the "data-value" problem—from which the value of datapoints derives—is shown to be equivalent to the linear-programming dual of the data-use problem.

Duality methods have been used to study the information-design problem by Kolotilin (2018), Galperti and Perego (2018), Dworczak and Martini (2019), Dworczak and Kolotilin (2019), and Dizdar and Kováč (2020). Our work differs from these papers in two important ways. First, they exploit the dual as a tool to compute and study the solution to the original primal problem. We instead put the dual at the center of our analysis and use it to address an independent economic question—namely, what is the value of data?—which is of interest irrespective of the primal problem. Second, unlike these papers, we do not focus on single-receiver Bayesian persuasion problems or employ a "belief approach" (as in Kamenica and Gentzkow, 2011). Rather, we study general information-design problems with multiple agents interacting strategically through the notion of Bayes-correlated equilibrium. This connects our work to an earlier literature on the characterization of correlated equilibria in complete-information games, which includes Nau and McCardle (1990), Nau (1992), and Myerson (1997). Finally, duality methods have also been applied in the mechanism-design literature, at least since Myerson (1983) and Myerson (1984), and more recently to study and characterize informationally robust mechanisms (e.g., Du, 2018; Brooks and Du, 2020, 2021).

Our paper contributes to bridging the literature on information design and the growing literature on data markets (see Bergemann and Bonatti, 2019, for a review). Closest to our paper are Bergemann and Bonatti (2015) and Bergemann et al. (2018), who build on earlier contributions of Admati and Pfleiderer (1986, 1990) to study markets where a buyer purchases an information product and uses its signal realization to solve a decision problem. This involves two key differences from our approach. First, their information product is a statistical experiment—not its signal realizations, which are assumed to be not contractible. This implies that values and prices are assigned only to those products ex ante, before any signal realization. By contrast, our approach can be viewed as assigning values to the signal realizations ex post. The second difference is that our potential buyer of data (i.e., our principal) uses it not to solve a decision problem, but to mediate strategic interactions. While assigning ex-post values would be standard in the settings of Bergemann and Bonatti (2015) and Bergemann et al. (2018), it is a novel rich problem in our case. Finally, data markets also raise important questions about privacy (see Acquisti et al., 2016, for a review). Acemoglu et al. (2021) and Bergemann et al. (2021) examine the externalities and market distortions that occur between agents who supply to a common intermediary private data which is correlated between them. These externalities are conceptually distinct from the ones we highlight. Ali et al. (2020) examine when giving consumers control over their private data can help them benefit from personalized pricing. Our results provide a benchmark to understand how different privacy regulations could affect the value of consumers' data for its users. We explore this question in a related project.

2 Model

Primitives. Let $i = 0$ denote the principal (she). Let $I = \{1, \dots, n\}$ denote a set of agents (he). For $i \in I$, let A_i be the finite set of actions under agent i's control. Let $A = A_1 \times \cdots \times A_n$. Let Ω be a finite set with typical element ω. Let $u_i : A \times \Omega \to \mathbb{R}$ be the utility function of party $i = 0, \dots, n$. Let $\Gamma_\omega = \{I, (A_i, u_i(\cdot, \omega))_{i=0}^n\}$. We will refer to Γ_ω as an *interaction* of type ω. If ω were commonly known, each interaction would define a complete-information game. However, only the principal observes ω. Let $q \in \mathbb{R}_+^\Omega \setminus \{0\}$ denote a collection of interactions, where $q(\omega)$ is the quantity of interactions of type ω. The primitive set of possible interactions $\Gamma = \{\Gamma_\omega\}_{\omega \in \Omega}$ and q are common knowledge.

Interpretation. We interpret Γ and q as defining the following mediation problem for the principal. An interaction defines the basic unit of the problem. The vector q defines the collection of physical, distinct, and independent interactions in the problem, which are drawn from some underlying infinite population characterized by Γ. The principal has exclusive access to all the variables that characterize each interaction, which we hereafter call its data for short and summarize with the type ω. This data can be interpreted—and modeled (see below)—as measurements or observations about some underlying, more primitive, payoff-relevant variables. For now, assume such measurements fully reveal these variables to the principal. Since each interaction and its data are an instance of some type ω from the Γ population, we will refer to the interaction-data pair as an *ω-datapoint* and to their collection q as the principal's *database*.[6] We can visualize this as a spreadsheet where rows are labeled by the interactions and columns by their characterizing variables, so each entire row is a datapoint. Using her data, the principal mediates each interaction by providing information about it to the involved n agents (as we formalize shortly). The agents use this information and the commonly known Γ and q to form beliefs about which type of interaction they face and hence to choose their actions. Thus, our principal is similar to the omniscient information designer in Bergemann and Morris (2019).

Information-Design Problem. The principal publicly commits to a mediation mechanism which takes the form of a information structure that, for each datapoint, produces a private signal about ω for each agent i aimed at influencing his choice of a_i. By standard arguments (Myerson, 1983, 1984; Bergemann and Morris, 2016), we can focus on mechanisms in which signals take the form of recommendations about which action to choose, subject to obedience: Each agent finds it optimal to follow his recommendations. A mechanism is then a function $x : \Omega \to \Delta(A)$, where $x(a|\omega)$ can be interpreted as the share of the quantity $q(\omega)$ of ω-datapoints that lead to the recommendation profile a. Formally, the information-design problem is

$$\mathcal{U}_q : \qquad \max_x \sum_{\omega\in\Omega, a\in A} u_0(a,\omega)x(a|\omega)q(\omega)$$
$$\text{s.t. for all } i \in I \text{ and } a_i, a_i' \in A_i,$$
$$\sum_{\omega\in\Omega, a_{-i}\in A_{-i}} \left(u_i(a_i, a_{-i}, \omega) - u_i(a_i', a_{-i}, \omega)\right) x(a_i, a_{-i}|\omega)q(\omega) \geq 0. \qquad (1)$$

Each constraint (1) is equivalent to requiring that a_i maximizes agent i's expected utility conditional on the information conveyed by a_i given x and the database q.

We assume that $\mathcal{U}_q$ satisfies the following minor regularity property, which holds generically in the space of agents' payoff functions.[7]

Assumption 1 (Non-degeneracy). *Of the constraints (1) that define the feasible set of mechanisms x for $\mathcal{U}_q$, no more than $|A \times \Omega|$ are ever active at the same time.*

[6] The assumption of a continuum of datapoints seems a reasonable approximation in a world of "big data."

[7] For more details, see Remark 1 in Appendix B.

Hereafter, denote any optimal mechanism by x_q^*, which exists by standard arguments, and let

$$U^*(q) = \sum_{\omega \in \Omega, a \in A} u_0(a, \omega) x_q^*(a|\omega) q(\omega).$$

A special case of this problem arises when all parties have perfectly aligned interests (i.e., u_i is an affine transformation of u_0 for all $i = 1, \ldots, n$). In this case, the obedience constraints (1) never bind and can be omitted. Thus, effectively the principal faces a collection of distinct decisions, one for each datapoint. For this reason, hereafter we will simply refer to this important benchmark with the term *decision problem*. When instead interests are not perfectly aligned, obedience constraints may affect the solution of the problem. We shall refer to this class of problems as *mediation problems*.

Our Goal. In the problem above, the principal optimally combines inputs (i.e., datapoints) to produce outputs (i.e., recommendations). In this paper, we use duality methods to pin down the value of each datapoint for the principal. We then characterize the properties of these values and study the principal's preferences over different databases.

2.1 Examples

Our general framework allows for many applications. We outline a couple of examples.

E-commerce. In our motivating example from the introduction, the platform is the principal and the basic unit of her mediation problem is an interaction between a buyer and the monopolist seller. The data of each interaction consists of the platform's records about the corresponding buyer, which may reveal—perhaps imperfectly—his valuation for the seller's product and thus pin down the interaction type ω. Thus, a datapoint is the triple composed by the seller, the buyer, and the record about the buyer.[8] The fact that the buyer is fully informed about ω can be reconciled with the assumption that only the principal has access to the data: For the purpose of the analysis, we can fold the buyer's behavior into the payoff function of the uninformed seller and treat him as the only agent (as in Bergemann et al., 2015). This setting can be generalized to allow for competition among possibly differentiated sellers (as in Elliott et al., 2020). In this case, each interaction involves one buyer and multiple sellers, each producing a single product and offering a price based on the signal about the buyer conveyed by the platform. Each buyer's record reveals his valuation for each seller's product in each interaction. We can again treat the sellers as the only agents. △

[8] Even though the seller is the same in all interactions here, it is conceptually useful to include it as part of the datapoint.

Routing Games. A navigation app uses data about routes' conditions to direct traffic by providing drivers with information—for instance, in the form of recommended paths and travel times. Das et al. (2017) propose a simple way to model this problem, which is of course much more complex in reality. Suppose the app (principal) seeks to minimize congestion. We can think of an interaction as consisting of a group of agents who simultaneously choose, say, one of two paths between some origin and destination. The travel time for one path is certain and increasing in how many agents choose it; for the second, it is independent of how many agents choose that path, but depends on some uncertain event (e.g., construction work). Only the app knows this as part of its data. For simplicity, suppose each such interaction happens in a different city so that they are independent in all respects, including the distribution of the uncertain event. A datapoint then consists of an interaction (agents and paths) and its data (what event realized). A database is the collection of such datapoints across the cities served by the app. If this collection is large, its composition q should reflect the primitive distribution of the uncertain event (i.e., the probability of construction on a given path). $\triangle$

2.2 Discussion of the Model

Data as Measurements. For each mediated interaction, its data is ultimately a signal about some underlying, more primitive, payoff-relevant variables. Formally, let Θ be some finite set and $\phi \in \Delta(\Theta)$. For every $i = 0, \ldots, n$, let $\hat{u}_i : A \times \Theta \to \mathbb{R}$ be the primitive payoff function of party i. It will be convenient to describe the signals about θ in terms of partitions (following Gentzkow and Kamenica, 2016). Let ζ be a random variable that is independent of θ and uniformly distributed on $[0, 1]$. Let a signal structure be a finite partition Ω of $\Theta \times [0, 1]$ such that each cell $\omega \in \Omega$ is a non-empty measurable subset of $\Theta \times [0, 1]$. For each $\omega \in \Omega$, the probability of ω conditional on θ is equal to $\lambda(\{\zeta : (\theta, \zeta) \in \omega\})$, where λ is the Lebesgue measure on $[0, 1]$. The interpretation is that each ω is a possible signal realization (i.e., piece of data) that the principal observes about θ. Given any such Ω, we can define the induced utility function as $u_i(a, \omega) = \mathbb{E}_\Omega[\hat{u}_i(a, \theta)|\omega]$.

This framework allows us to capture databases for which the principal has precise measurements about some interactions and imprecise measurements about others. Thus, in this paper the colloquial expression "having more data" can mean two things: (1) the principal has more data about *one* interaction in the sense of observing a more informative measurement ω about that interaction; (2) the principal has a larger database with more datapoints (i.e., interactions and their data). We will analyze both. Note that even if ω is fully uninformative about θ (like ω° in the example), being the only one to know this for some interaction can still grant the principal an informational advantage.

Simultaneous vs Sequential Mediation. We interpret the principal as mediating all interactions in her database simultaneously. For instance, in the illustrative example, the platform may mediate all buyer-seller trades simultaneously, where the seller sets one price for each market segment identified by the signal that the platform sends. An equally valid interpretation is that the principal commits to a mechanism for the whole database and then interactions are drawn independently and mediated one at a time. In our example, the platform may mediate the trade between a buyer and the seller one at a time, sending signals according to its mediation plan. Depending on the application, one interpretation may fit better.

Simplifications. For ease of exposition, we simplified the model in several ways. Neither changes the analysis or its interpretations. First, we can allow the principal to also choose an action $a_0 \in A_0$ for each mediated interaction. In this case, a mechanism x would also have to specify a_0 for each ω. Second, we can allow each agent i to also observe privately some own data about the interaction he is in. We can again model such data as a partitional signal structure Ω_i, as we did above for the principal. The interpretation is that each ω_i is a piece of data, or measurement, that i can observe about the underlying θ. For example, in the case of a platform mediating trade between a buyer and possibly multiple sellers, ω_i can refer to the quality or the history of customer reviews of seller i's product. Let $\Omega = (\Omega_0, \ldots, \Omega_n)$ be a vector of partitions for all parties, with typical element $\omega = (\omega_0, \ldots, \omega_n)$. Now the whole vector ω summarizes the data characterizing an interaction and defines its type. The key assumption is that the principal also observes the private data of each agent—i.e., the entire $\omega = (\omega_0, \ldots, \omega_n)$—as does the omniscient designer in Bergemann and Morris (2019). However, now agent i may know something about what the principal knows regarding the interaction i is in. We will explain in Appendix B how to adapt the analysis to take this into account (see proof of Lemma 1).[9]

Two Comments on Terminology. First, our Ω is akin to what is often called the set of states of the world. Indeed, we can view an interaction as being characterized by its realized state. We do not use this terminology to avoid confusion: In our physical interpretation, a database contains multiple interactions each with its state, so multiple states are simultaneously realized and q need not be a probability measure. Second, each Ω_i is akin to what is often called the set of party i's types. We can then view an interaction as being characterized by the profile of all its participants' types. We instead use "type" to refer to the whole profile because our analysis

[9] The case in which the principal has to elicit the agents' private data significantly changes the principal's problem. We consider this case in an ongoing separate project.

focuses on how the principal mediates interactions that differ in their characterizing data *as a whole*. This use is also consistent with viewing ω_0 as the principal's type when she alone observes data, as this type is defined by what data characterizes the interaction she is mediating.

3 The Value of Datapoints

We begin by illustrating how, for any information-design problem, we can characterize the value that the principal derives from each datapoint. Our approach relies on a standard interpretation of a linear program as the problem of optimally using scarce resources (Dorfman et al., 1987, p. 39). We think of the information-design problem as a "data-use" problem: Each mechanism x combines scarce inputs—i.e., datapoints—to produce output—i.e., information in the form of recommendations for each agent. Following Dorfman et al. (1987, p. 39), we then exploit the dual of the data-use problem to assign a unit value to each datapoint in the database based on standard economic principles of opportunity cost and marginal considerations.

We refer to the problem of assigning values to datapoints as the *data-value* problem. Let $b = (b_1, \ldots, b_n)$ be a profile such that $b_i : A_i \to \mathbb{R}_{++}$ for all i and $\ell = (\ell_1, \ldots, \ell_n)$ be a profile such that $\ell_i : A_i \to \Delta(A_i)$ for all i.[10] Given (b, ℓ), for each $i \in I$ and (a, ω) define

$$t_i(a,\omega) = b_i(a_i) \sum_{a_i' \in A_i} \left(u_i(a_i, a_{-i}, \omega) - u_i(a_i', a_{-i}, \omega)\right) \ell_i(a_i'|a_i) \tag{2}$$

and $t(a,\omega) = \sum_{i \in I} t_i(a,\omega)$. The data-value problem is then

$$\begin{aligned} \mathcal{V}_q : \quad & \min_{v,b,\ell} \sum_{\omega \in \Omega} v(\omega) q(\omega) \\ & \text{s.t. for all } \omega \in \Omega, \\ & v(\omega) = \max_{a \in A} \left\{ u_0(a,\omega) + t(a,\omega) \right\}, \end{aligned} \tag{3}$$

Hereafter, we will denote any optimal solution of $\mathcal{V}_q$ by (v_q^*, b_q^*, ℓ_q^*) and the induced functions t by t_q^*. Note that, by standard linear-programming arguments, (v_q^*, b_q^*, ℓ_q^*) is unique generically with respect to q: The set of databases for which it is not unique has measure zero.

We refer to equation (3) as the *value formula*, which defines the main object of interest in this paper. The reason hinges on the following relation between the data-value problem $\mathcal{V}_q$ and the data-use problem $\mathcal{U}_q$ and on the ensuing interpretation.

[10] We explain how to interpret (b, ℓ) in Section 5.

Lemma 1. *For any database q, $\mathcal{V}_q$ is equivalent to the dual of $\mathcal{U}_q$. Thus, for every x_q^* and (v_q^*, b_q^*, ℓ_q^*)*

$$\sum_{\omega \in \Omega} v_q^*(\omega) q(\omega) = U^*(q) \tag{4}$$

All proofs are in the Appendix B. The duality relation between $\mathcal{V}_q$ and $\mathcal{U}_q$ follows from basic linear-programming results. Yet, applied to our specific problem, it shows that $v_q^*(\omega)$ captures the effect of a marginal change in $q(\omega)$ on the principal's total payoff. More precisely, in $\mathcal{U}_q$, by choosing x, the principal effectively chooses a joint measure $\chi \in \mathbb{R}_+^{\Omega \times A}$, which must satisfy $\sum_{a \in A} \chi(a, \omega) = q(\omega)$; that is, the use of ω-datapoints to produce recommended action profiles a must exhaust their stock $q(\omega)$ in the database. Formally, $v(\omega)$ is the multiplier of this constraint. Following classic interpretations (Gale, 1989, p. 12), we will refer to $v_q^*(\omega)$ as the *unit value* of ω-datapoints. It is worth noting that v_q^* can depend on q for mediation problems (hence, the subscript). By contrast, for decision problems $v_q^*(\omega)$ equals $\max_{a \in A} u_0(a, \omega)$ for all ω and hence is independent of q.

In the next subsections, we characterize the properties of v_q^*, including its applicability to non-local changes in q. Before that, we present three ways in which the value of datapoints could be interpreted and used in practice.

Willingness To Pay. Consider the problem of a principal who can add new datapoints to her database. Suppose she is offered one datapoint. At this stage, she may immediately observe its characterizing data and so ω, or she may only have a belief about what this data is. Let this belief be $\rho \in \Delta(\Omega)$. Either way, suppose that after buying the datapoint she observes ω. The values v_q^* are instrumental in determining the principal's willingness to pay (WTP) for this new datapoint. In fact, a multiplier of a resource constraint is also usually interpreted as the shadow price of that resource—hence, its user's willingness to pay for an additional unit of it. Moreover, adding only one datapoint to a database constitutes a marginal change that leaves q unaffected. Exploiting the generic uniqueness of v_q^*, we have that the expected marginal increase of the principal's total payoff from adding such a datapoint is

$$v_q^*(\rho) = \sum_{\omega} \rho(\omega) v_q^*(\omega),$$

where we extend $v_q^* : \Omega \to \mathbb{R}$ to $v_q^* : \Delta(\Omega) \to \mathbb{R}$. Thus, for any belief ρ the principal may have when offered new datapoints, we can interpret $v_q^*(\rho)$ as her WTP.

Individual Contribution. The quantity $v_q^*(\omega)$ not only captures the value of an additional ω-datapoint. As it applies linearly to all datapoints of the same type, it also provides an assessment of the contribution that each ω-datapoint individually makes to the total payoff $U^*(q)$. As such,

$v_q^*(\omega)$ represents a possible way to compensate ”owners” of datapoints. Paraphrasing Dorfman et al. (1987, p. 43), this interpretation is reminiscent of the operation of a competitive market in which the principal is forced by competition to offer the “owners” of each type of datapoints the full value to which their inputs give rise, while competition among these “owners” drives down this value to the minimum consistent with this limitation. Gale (1989, Chapter 3.5) also shows how dual problems deliver competitive prices of scarce inputs.

Value of Information. Each datapoint generates value for the principal in possibly multiple and complex ways. As argued in Section 2, a datapoint identifies an interaction together with its data. The value of a datapoint then depends on (a) the fact that the principal can mediate the interaction and (b) the *information* contained in its data. This information allows the principal to know her payoff and the agents' incentives, thereby determining the informational advantage that she can exploit in mediating this as well as *other* interactions. A natural question is whether we can isolate (b), to which we may refer as the value of the information contained in a datapoint for mediation problems. Intuitively, one way to do this is to “net out” from $v_q^*(\omega)$ the value the principal would derive if she did not have the data that characterizes that interaction, while still mediating it. Let ω° denote a datapoint involving completely uninformative observations about the interaction's relevant variables.[11] We may then quantify (b) as

$$v_q^*(\omega) - v_q^*(\omega^\circ), \qquad \omega \in \Omega. \tag{5}$$

It is immediate to see that this intuitive procedure involves some subtle intricacies. What the principal does with ω°-datapoints can depend on q and so does $v_q^*(\omega^\circ)$. This is because the principal remains the only one to know that she knows nothing about a specific interaction and how she can exploit this depends on the composition of the database. By contrast, for decision problems the net value $v_q^*(\omega) - v_q^*(\omega^\circ)$ is defined in absolute terms independently of q. For these reasons, quantifying the value of information for mediation problems requires to first characterize v_q^*. We view this as a feature, not a bug, of our approach. In fact, this relative value of information has a sound economic intuition. We will return to this value of information in Section 4.

We illustrate these three perspectives on the value of datapoints through our running example.

Example. (Cont'd) The database of our online platform contains three types of datapoints: $\omega \in \{\omega^\circ, 1, 2\}$. Let $h > \frac{1}{2}$ be the probability that for an ω°-datapoint the buyer has valuation 2. Table 1 reports v_q^*, which we calculate in Appendix C:

[11] Using the formalism of partitions of $\Theta \times [0,1]$, ω° is a cell of the form $\Theta \times L$, where L is a Lebesgue measurable subset of $[0,1]$.

		$v_q^*(\omega^\circ)$	$v_q^*(1)$	$v_q^*(2)$
(i)	$q(1) \in (0, (2h-1)q(\omega^\circ))$	0	$h/(2h-1)$	0
(ii)	$q(1) \in ((2h-1)q(\omega^\circ), q(2) + (2h-1)q(\omega^\circ))$	$1-h$	1	0
(iii)	$q(1) \in (q(2) + (2h-1)q(\omega^\circ), \infty)$	h	0	1

Table 1: Platform Example: Values of Datapoints

Given any q, we can interpret each $v_q^*(\omega)$ as the platform's WTP for adding ω-datapoints to the original database. For instance, if the original q has relatively few 1-datapoints (condition (i)), the WTP is $\frac{h}{2h-1} > 1$. As it adds more 1-datapoints, the WTP drops to 1 and eventually to 0. We will examine this property in Section 3.2.

Table 1 also illustrates the point that the value of the information contained in datapoints and the willingness to pay for learning (WTPL) depends on q. In particular, note that $hv_q^*(2) + (1-h)v_q^*(1) - v_q^*(\omega^\circ)$ is positive when q satisfies condition (i), but is zero otherwise. Note that this zero WTPL does not arise because the platform does not change how it uses its data as it learns the valuation of buyers in ω°-datapoints. We will examine this property in Section 4.

Finally, with regard to compensating people for their data, imagine that the ω°-datapoints correspond to buyers who blocked the platform's tracking of their behavior. If they now allow tracking, how should the resulting increase in the database value be allocated among them? For illustration, suppose the initial q and final q' satisfy condition (ii) and $U^*(q') - U^*(q) > 0$. It seems intuitive that this should be allocated taking into account what data each buyer shares. Since all the extra value comes from buyers with valuation 1 (because $v_q^*(2) = v_{q'}^*(2) = 0$), one solution is to divide it equally only among them and give nothing to buyers with valuation 2. $\triangle$

Before continuing with the analysis, it is useful to show that the value of each datapoint has a lower bound defined by the primitives Γ. For every $\omega \in \Omega$, let $CE(\Gamma_\omega)$ be the set of correlated equilibria of the complete-information game Γ_ω. We then obtain the following.[12]

Lemma 2 (Lower Bound). *For every database q,*

$$v_q^*(\omega) \geq \bar{u}(\omega) := \max_{y \in CE(\Gamma_\omega)} \sum_{a \in A} u_0(a, \omega) y(a), \qquad \omega \in \Omega.$$

As in Maskin and Tirole (1990), we say that x is *full-information incentive compatible* (FIC) if $x(\cdot|\omega) \in CE(\Gamma_\omega)$ for all $\omega \in \Omega$. We say that $\mathcal{U}_q$ is a *full-disclosure problem* if there is an optimal x_q^* that is FIC. Lemma 1 and 2 imply that this is the case if and only if $v_q^*(\omega) = \bar{u}(\omega)$ for all ω.

[12]Note that this property is not directly implied by the obvious fact that $U^*(q) \geq \sum_{\omega \in \Omega} \bar{u}(\omega) q(\omega)$.

3.1 Value Decomposition and Data Externalities

We now delve deeper into what determines the value of datapoints. Key to this will be comparing v_q^* with the payoff the principal directly derives from the outcomes implemented at each datapoint. Given any optimal x_q^*, we define the *direct payoff* of ω-datapoints as

$$u_q^*(\omega) = \sum_{a \in A} u_0(a,\omega) x_q^*(a|\omega), \quad \omega \in \Omega.$$

As anticipated in our example from the introduction, u_q^* can differ from v_q^*. This difference arises because u_q^* ignores that the principal may be able to implement $x_q^*(\cdot|\omega)$ with ω-datapoints only as a result of having other ω'-datapoints in her database. The gap between the direct payoff of ω-datapoints, $u_q^*(\omega)$, and their actual value, $v_q^*(\omega)$, quantifies an externality that captures the effects an ω-datapoint exerts on the direct payoff of other datapoints—of the same or a different type. In this section, we characterize this externality and argue it is a defining feature of mediation problems.

Proposition 1. *For all $\omega \in \Omega$, $v_q^*(\omega) = u_q^*(\omega) + t_q^*(\omega)$ where*

$$t_q^*(\omega) := \sum_{a \in A} t_q^*(a,\omega) x_q^*(a|\omega) \overset{a.e.}{=} \sum_{\omega' \in \Omega} \frac{\partial u_q^*(\omega')}{\partial q(\omega)} q(\omega') \tag{6}$$

This result highlights several aspects of the value of datapoints, which we now discuss in order.

First, the value of each datapoint can be decomposed into two parts: The direct payoff $u_q^*(\omega)$ of each ω-datapoint and a valuation, $t_q^*(\omega)$, of the indirect effects that it generates. This decomposition clarifies more formally why u_q^* is ill-suited to assessing the actual value that each datapoint generates for the principal.

A second aspect is that the indirect effects captured by $t_q^*(\omega)$ are akin to an externality exerted by ω-datapoints on all other datapoints. Indeed, the last part of (6) shows that $t_q^*(\omega)$ captures the marginal effect of an ω-datapoint on the direct payoff of *all* datapoints, including those of the same type. This externality is purely informational in nature, as $\frac{\partial}{\partial q(\omega)} u_q^*(\omega') = \sum_a u_0(a,\omega') \frac{\partial}{\partial q(\omega)} x_q^*(a|\omega')$. In other words, each ω-datapoint can generate an externality on other datapoints because, by being in the database, it can affect the principal's informational advantage and hence the optimal way x_q^* in which the whole database is used. This adjustment in x_q^* can arise because changing $q(\omega)$ can render x_q^* no longer feasible or no longer optimal.

A natural question is which datapoints generate a positive externality and which a negative externality. Combining Lemma 1 and 2 and Proposition 1, we obtain the following.[13]

[13]Lemma 1 implies $\sum_{\omega \in \Omega} t_q^*(\omega) q(\omega) = 0$. Lemma 2 and Proposition 1 imply $t_q^*(\omega) \geq \bar{u}(\omega) - u_q^*(\omega)$ for all ω.

Corollary 1. *$t_q^*(\omega) < 0$ for some ω if and only if $t_q^*(\omega') > 0$ for some ω'. Moreover, $t_q^*(\omega) < 0$ implies $u_q^*(\omega) > \bar{u}(\omega)$, while $u_q^*(\omega) < \bar{u}(\omega)$ implies $t_q^*(\omega) > 0$.*

The first part shows that, whenever there are some externalities, there is cross-subsidization of value from datapoints with $t_q^* < 0$ to datapoints with $t_q^* > 0$. The latter's value v_q^* exceeds their direct payoff and hence they must extract this extra value from datapoints with $t_q^* < 0$. The second part of the corollary explains this cross-subsidization. Datapoints with $t_q^* < 0$ generate a direct payoff that exceeds the full-information payoff $\bar{u}$. Note that $u_q^*(\omega) > \bar{u}(\omega)$ requires that there be a such that $x_q^*(a|\omega) > 0$ and $u_0(a,\omega) > \bar{u}(\omega)$. That is, the principal manages to achieve a payoff with ω-datapoints that would never be possible by fully disclosing such datapoints and hence relies on pooling ω-datapoints with other ω'-datapoints in the database. This help from ω'-datapoints justifies why $t_q^*(\omega') > 0$ and their value exceeds their direct payoff. Conversely, if $t_q^*(\omega) < 0$, an ω-datapoint benefits from externalities caused by other datapoints and hence has to "repay" them, which lowers its value. The corollary also offers a sufficient condition for $t_q^* \neq 0$ that is simple to check, but may not always apply. One can easily construct examples where $u_q^*(\omega) \geq \bar{u}(\omega)$ for all ω and $t_q^* \neq 0$.[14] These cases are covered by a condition presented later in Appendix A.[15]

A third aspect highlighted by Proposition 1 is that the externalities in terms of direct payoffs are tightly related to how the principal exploits the agents' primitive incentives. By (6), we can view $t_q^*(\omega)$ as an aggregation of externalities that ω-datapoints generate by inducing specific actions a. These are inversely related to the payoff the principal gets, in the following sense.[16]

Corollary 2. *Suppose $x_q^*(a|\omega) > 0$ and $x_q^*(a'|\omega) > 0$. Then, $u_0(a,\omega) > u_0(a',\omega)$ if and only if $t_q^*(a,\omega) < t_q^*(a',\omega)$.*

Thus, inducing actions whose payoff exceeds $\bar{u}$ by more, for instance, requires paying larger externalities to other datapoints. Since $t_q^*(a,\omega) = \sum_{i\in I} t_{q,i}^*(a,\omega)$, we can view $t_{q,i}^*(a,\omega)$ as how much agent i contributes to the externality. Recall that $t_{q,i}^*(a,\omega)$ differs from zero only if $\ell_{q,i}^*(a_i'|a_i) > 0$ for some a_i' (see (2)). By standard arguments (complementary slackness),

[14] For instance, this property holds for the leading judge-prosecutor example in Kamenica and Gentzkow (2011).

[15] One may wonder whether t_q^* is related to the payoff difference $\bar{u}(\omega) - u_q^*(\omega)$. While it can happen that $t_q^*(\omega) = \bar{u}(\omega) - u_q^*(\omega)$ for some ω, there is only one trivial case where this holds for all ω. In this case, we must have $v_q^*(\omega) = \bar{u}(\omega)$ for all ω, which by Proposition 4 implies that $t_q^* \equiv 0$ for all q.

[16] This follows from complementary slackness, which in our case says that $x_q^*(a|\omega) > 0$ implies $v_q^*(\omega) = u_0(a,\omega) + t_q^*(a,\omega)$.

$\ell^*_{q,i}(a'_i|a_i) > 0$ only if

$$\sum_{\omega,a_{-i}} \Big(u_i(a_i,a_{-i},\omega) - u_i(a'_i,a_{-i},\omega)\Big) x^*_q(a_i,a_{-i}|\omega)q(\omega) = 0; \tag{7}$$

the converse also holds generically in q. In words, $\ell^*_{q,i}(a'_i|a_i) > 0$ if and only if agent i is indifferent between a_i and a'_i conditional on receiving recommendation a_i from x^*_q. This fact implies the following.

Corollary 3. *The agents who contribute to the externality $t^*_q(\omega)$ are only those whom x^*_q renders indifferent with the actions it recommends using ω-datapoints (i.e., (7) holds).*

Note that this result differs from the immediate fact that optimal solutions of linear programs occur on the boundary of the feasible set, which here means that some obedience constraint must bind. As q varies, x^*_q and hence $t^*_q(\omega)$ may change. However, as long as b^*_q and ℓ^*_q do not change (see Proposition 2 below), how each agent contributes to $t^*_q(\omega)$ does not change. Section 5 explains further how the principal exploits the agents to determine their contributions to these externalities.

Our externalities *between* datapoints differ from other data-driven externalities studied in the literature. One distinction is that they never arise for decision problems—and, in fact, for full-disclosure problems as implied by Proposition 4 below. For both such problems $v^*_q(\omega) = u^*_q(\omega) = \bar{u}(\omega)$ and therefore $t^*_q(\omega) = 0$ for all $\omega \in \Omega$ and q.[17] In Acemoglu et al. (2021) and Bergemann et al. (2021), externalities arise due to correlation between the data of different consumers, which in our framework would correspond to externalities *within* a single datapoint between its dimensions. These externalities can also arise in decision problems. Calzolari and Pavan (2006) analyze information externalities between sequential interactions. We do not consider such interactions.

3.2 Scarcity Principle and Preferences over Databases

In this section we study how the value of datapoints depends on the database composition, which in turn characterizes the principal's preferences over databases. We first show that v^*_q is constant with respect to local, yet discrete, changes in q and hence its meaning goes beyond that of a *marginal* value.

[17]The converse is not true: It is possible to construct examples where $t^*_q(\omega) = 0$ and $v^*_q(\omega) > \bar{u}(\omega)$ for all ω.

Proposition 2 (Stability)**.** *There exists a finite collection $\{Q_1, \ldots, Q_K\}$ of open, convex, and disjoint subsets of $\mathbb{R}^\Omega_+$ such that $\cup_k Q_k$ has full measure and, for every k, (v^*_q, b^*_q, ℓ^*_q) is unique and constant for $q \in Q_k$.*

In fact, each Q_k is the interior of a cone in the space of databases $\mathbb{R}^\Omega_+$.[18] Importantly, v^*_q is constant even though the principal may adjust how she uses her data when q changes. Indeed, we can show (see Remark 1 in Appendix B) that within each cone, while $v^*_q(\omega)$ is constant, the optimal $x^*_q(\omega)$ changes as a function of q. Intuitively, this is because x^*_q has to be fine-tuned to maximally exploit the agents' incentives. By contrast, v^*_q depends only on which agents' incentives are exploited, but not by how much. Note that for decision problems $K = 1$ since v^*_q is independent of q.

Next, we analyze how v^*_q varies as a function of general changes to q. For example, do datapoints become more valuable as they become scarcer? The next result establishes a general "scarcity principle" for datapoints. To this end, for every q define the implied share μ_q of datapoint of each type by

$$\mu_q(\omega) = \frac{q(\omega)}{\sum_{\omega'} q(\omega')}, \qquad \omega \in \Omega.$$

Proposition 3 (Scarcity Principle)**.** *Consider databases q and q'. If $\mu_q(\omega) > \mu_{q'}(\omega)$, then $v^*_q(\omega) \leq v^*_{q'}(\omega)$. Moreover, for every $\omega \in \Omega$ there exists $\bar{\mu}(\omega) < 1$ such that, if $\mu_q(\omega) > \bar{\mu}(\omega)$, then $v^*_q(\omega) = \bar{u}(\omega)$.*

In particular, this implies that $v^*_q(\omega)$ is weakly decreasing in $q(\omega)$—for any selection from the optimal solution correspondence of $\mathcal{V}_q$. It is perhaps surprising that this property holds for *any* information-design problem. Different types of datapoints can contribute to the principal's informational advantage in different ways; yet, datapoints always becomes more valuable as they become scarcer. This implies that, holding fixed the quantity of all other types of datapoints, the principal's demand for ω-datapoints is downward sloping and converges to the lower bound $\bar{u}(\omega)$ when $q(\omega)$ is sufficiently large. Note that if $\mu_q(\omega) > \mu_{q'}(\omega)$, then we must have $\mu_q(\omega') < \mu_{q'}(\omega')$ for some ω' and therefore $v^*_q(\omega') \geq v^*_{q'}(\omega')$. For decision problems v^* is independent of q and therefore this scarcity principle holds trivially.

With regard to the principal's overall preferences over databases, it is easy to see that they are always convex for any mediation problem. Indeed, her total payoff $U^*(q)$ is a concave

[18] It is easy to see that the value of a datapoint is constant along the rays in the space of databases: If $q' = \alpha q$ for $\alpha > 0$, then $v^*_q = v^*_{q'}$. This is because the relative frequency of datapoints in q and q' is the same, and this is the only thing that matters to determine agents' incentives.

function of the database q; hence, her iso-payoff curves—i.e., the set of qs that lead to the same value of U^*—define convex upper-contour sets in $\mathbb{R}^{\Omega}_{+}$.[19] These iso-payoff curves are entirely characterized by v^*_q, as the marginal rate of substitution between ω- and ω'-datapoints is generically $v^*_q(\omega)/v^*_q(\omega')$.

For all mediation problems, choosing an optimal database subject to a budget constraint is then a well-behaved problem. Suppose the principal can acquire datapoints of each type ω at a given market price $p(\omega) > 0$.[20] The optimal q is characterized by

$$\max_{v \in v^*_q} \frac{v(\omega)}{v(\omega')} \geq \frac{p(\omega)}{p(\omega')} \geq \min_{v \in v^*_q} \frac{v(\omega)}{v(\omega')}, \qquad \omega, \omega' \in \Omega,$$

where, with slight abuse of notation, v^*_q stands for the *set* of optimal solutions at q. This is equivalent to $p \in \partial U^*(q)$, where $\partial U^*(q)$ is the superdifferential of U^* at q. In this way, we can use v^*_q to characterize the principal's *demand functions* for datapoints, thus enabling a general study of the demand side of the "market for data." The actual price paid for datapoints in this market is of course determined by the interplay of demand and supply. Under perfect competition, Dorfman et al. (1987) and Gale (1989) provide arguments for equality between v^*_q and equilibrium prices.

Convexity of iso-payoff curves and, hence, imperfect substitutability between datapoints are hallmarks of non-trivial mediation problems. In a decision problem, the principal trivially discloses all datapoints, regardless of q. Therefore, $v^*_q(\omega) = \bar{u}(\omega)$ for all ω and q. In this case, iso-payoff curves are always hyperplanes (the analog of straight lines in two dimensions); hence, datapoints are perfectly substitutes. More generally, it is possible that in a mediation problem there is a particular database composition q at which it is optimal to fully disclose. In other words, the principal faces a *full-disclosure problem* at q. The next result shows that, when this is the case, the principal faces a full-disclosure problem at all q.

Proposition 4. *Fix Γ. Suppose $\mathcal{U}_q$ is a full-disclosure problem for some database $q \in \mathbb{R}^{\Omega}_{++}$. Then, $v^*_{q'}(\omega) = \bar{u}(\omega)$ for all ω all $q' \in \mathbb{R}^{\Omega}_{+}$ and, thus, $\mathcal{U}_{q'}$ is a full-disclosure problem for all q'.*

Requiring $q \in \mathbb{R}^{\Omega}_{++}$ is without loss of generality: If $q(\omega) = 0$ for some $\omega \in \Omega$, we can drop such ω to obtain Ω' and the same result holds with $q' \in \mathbb{R}^{\Omega'}_{++}$. Entirely removing some types

[19]Concavity follows because, through (4), we can view U^* as the minimization of a family of linear functions in the "parameter" q (see, e.g., Theorem 5.5 in Rockafellar, 1970). It is related directly to the concavification results in Mathevet et al. (2020) and indirectly to the individual-sufficiency results in Bergemann and Morris (2016).

[20]Note that in the special case with a unitary budget and $p(\omega) = 1$ for all ω, choosing q is isomorphic to choosing an optimal prior in $\Delta(\Omega)$.

of datapoints from a database changes the primitive Γ and can obviously render it optimal to fully reveal all the remaining data.

Proposition 4 has several implications. First, datapoints are perfect substitutes if and only if it is always optimal to fully disclose all datapoints. This is the only case where iso-payoff curves can be hyperplanes, which are entirely pinned down by the vector $\bar{u}$. Conversely, datapoints are imperfect substitutes or even complements if and only if it is never optimal to fully disclose all datapoints. In this case, iso-payoff curves are never hyperplanes and $v_q^* \neq \bar{u}$ for all $q \in \mathbb{R}_{++}^{\Omega}$. Second, we can identify whether a principal faces a non-trivial mediation problem or (essentially) a decision problem by detecting any imperfect substitutability between datapoints in her preference over databases. Third, showing that it is not optimal to fully disclose some ω for some $q \in \mathbb{R}_{++}^{\Omega}$ suffices to show that it is never optimal to fully disclose for all $q \in \mathbb{R}_{++}^{\Omega}$. More generally, Γ is all that one needs to know to establish the (sub)optimality of fully disclosing all data and that the principal faces a non-trivial mediation problem.[21]

4 The Value of Information in Mediation Problems

In this section, we consider the problem of acquiring more data in the sense of refining an existing datapoint by obtaining more informative measurements about this interaction. A key observation is that such a refinement transforms the original datapoint into a datapoint of a different type. Therefore, acquiring more informative data is equivalent to changing a database q into another q' by refining some of its original datapoints. We can then assess how much the principal *values information* by how she values refining q into q' according to her preference over databases, which we discussed in the previous section. Recall that in our leading example the platform may initially have no record of observations about a buyer who just joined its userbase (i.e., $\omega = \omega^\circ$). After several observations, it may learn that his valuation is high (i.e., $\omega = 2$). As a result, its database loses an ω°-datapoint and gains a 2-datapoint.

To study the value of information for mediation problems in this way, we only need to specify the space of possible databases appropriately. Intuitively, this space should contain all types of datapoints that correspond to the level of information the principal can conceivably have about an interaction. In the example, if each buyer's valuation is pinned down by age and gender, the types of datapoints correspond to the observation of neither variable, of only one variable, and of both variables for all their possible realizations.

[21] We provide sufficient conditions for this in terms of Γ in Appendix A.

We formalize this with the notion of a *complete space* of databases. Recall that we can view each ω as the cell of a finite partition of a more primitive space of uncertainty $\Theta \times [0,1]$ (Section 2.2). Thus, we can capture more informative data in terms of finer partitions. Let $\mathbf{\Omega} = \{\Omega^\circ, \Omega^1, \ldots, \Omega^K\}$ be a family of finite partitions of $\Theta \times [0,1]$ with the property that $\Omega^\circ = \{\omega^\circ\} = \Theta \times [0,1]$ and Ω^k is finer than Ω^{k-1} for all $k = 2, \ldots, K$. Given such an $\mathbf{\Omega}$, we say that the space of databases $\mathbb{R}_+^\Omega$ is complete if

$$\Omega = \bigcup_{k=1}^{K} \Omega^k.$$

That is, if the principal can refine ω-datapoints by acquiring better data, every type ω' of datapoints that can result from this refinement is taken into account. We include Ω° so that, if $q(\omega^\circ) > 0$ and $q(\omega) > 0$, the value of the information contained in ω-datapoints is well defined as given by $v_q^*(\omega) - v_q^*(\omega^\circ)$ (see (5)). Given a complete space, any acquisition of better data involves moving inside the space itself.

This formalism allows us to describe many forms of data acquisition, in terms of their level of informativeness and correlation across datapoints. We will start from a focal form that can be thought of as observing a signal for each datapoint to be refined, whose distribution is identical given ω, independent across datapoints, and equal to the composition of the underlying population of interactions. Concretely, for each ω°-datapoint the platform may observe its buyer's age or gender (or both) according to their distribution in the overall population of buyers. In this case, we have that each ω-datapoint changes to a refined ω'-datapoint with probability

$$\sigma_\omega(\omega') = \frac{\lambda(\{\zeta : (\theta, \zeta) \in \omega \cap \omega'\})}{\lambda(\{\zeta : (\theta, \zeta) \in \omega\})}, \qquad \omega' \in \operatorname{supp} \sigma_\omega,$$

where recall that λ is the Lebesgue probability on $[0,1]$. Note that $\operatorname{supp} \sigma_\omega$ can be a strict subset of $\{\omega' : \omega' \subseteq \omega\}$. Because we imagine databases with a large quantity—formally, a continuum—of datapoints, these transition probabilities allow us to describe the effects of acquiring better data about a share $\alpha \in [0,1]$ of ω-datapoints. By the Law of Large Numbers, $\sigma_\omega(\omega')$ is also the fraction of these datapoints that become ω'-datapoints. Given the initial quantity $q(\omega)$ of ω-datapoints, the new database q_α will contain the quantity $q_\alpha(\omega) = (1 - \alpha)q(\omega)$ of ω-datapoints and $q_\alpha(\omega') = q(\omega') + \alpha\sigma_\omega(\omega')q(\omega)$ of ω'-datapoints for each ω' such that $\sigma_\omega(\omega') > 0$. We will refer to (α, σ_ω) as an i.i.d. refinement. Note that there is no uncertainty about the composition q_α of the new database, even though there is uncertainty about which ω-datapoints become ω'-datapoints. In our example, given q if the seller knows that the platform has acquired better data about ω°-datapoints according to $(\alpha, \sigma_{\omega^\circ})$, he knows the resulting q_α.

Refining datapoints improves what the principal knows about them, but also changes her informational advantage. Thus, whether it increases the value of refined datapoints and ultimately benefits the principal is unclear. For clarity, we address these two aspects in separate results, starting from the value of datapoints.

Proposition 5. *Fix q, $\omega \in \Omega$, and i.i.d. refinements (α, σ_ω). In expectation, refined datapoints are weakly more valuable for all $\alpha \in [0,1]$, i.e.,*

$$\sum_{\omega' \in \Omega} v^*_{q_\alpha}(\omega')\sigma_\omega(\omega') - v^*_q(\omega) \geq 0.$$

Moreover, this expected gain equals zero for all α if there exists $a \in supp\, x^*_q(\cdot|\omega'')$ for $\omega'' = \omega$ and all $\omega'' \in supp\, \sigma_\omega$. The converse is true generically in the space of databases.

Refining a single ω-datapoint does not change q (as this is formally the same as $\alpha = 0$). Hence, the gain in value reflects only the fact that the principal is better informed about this datapoint. Yet, when $\alpha > 0$ and so $q_\alpha \neq q$, the resulting change in informational advantage can cause $v^*_{q_\alpha}$ to change as well. In particular, the quantity of the types of datapoints that result from the refinement increases, which reduces their value by the scarcity principle. Despite this negative effect, i.i.d. refinements always improve values in expectation. As shown shortly, this can fail for other forms of data refinement.

The result provides a sharp condition for the expected gain from acquiring better data about some mediated interactions to be *zero*. There must be a *common* action profile that the principal induces with positive probability for ω-datapoints as well as for every datapoint that ω-datapoints can turn into by the refinement. Importantly, one may think that refining datapoints leaves their value unaffected simply because the principal does not change how she uses them (i.e., x^*). This is incorrect: It is possible to construct examples where the expected value gain of refining ω-datapoints is zero, yet $x^*_{q_\alpha}(\cdot|\omega) \neq x^*_q(\cdot|\omega)$. This is in contrast to decision problems, where the gain from more information about a decision problem is positive if and only if it changes behavior.

Example. (Cont'd) Consider q that satisfies condition (ii) of Table 1. The corresponding x^*_q satisfies

$x^*_q(a\|\omega)$	$\omega = 1$	$\omega = 2$	$\omega = \omega^\circ$
$a = 1$	1	$\frac{q(1)-(2h-1)q(\omega^\circ)}{q(2)}$	1
$a = 2$	0	$1 - \frac{q(1)-(2h-1)q(\omega^\circ)}{q(2)}$	0

Table 2: Platform Example, x^*_q for Condition (ii) in Table 1.

Thus, as we vary q in the region defined by condition (ii), x_q^* changes while the expected value gain of refining ω°-datapoints remains zero. $\triangle$

We now turn to whether the principal benefits overall from data refinement.

Proposition 6. *Fix q, $\omega \in \Omega$, and i.i.d. refinements (α, σ_ω). The expected marginal effect of refining ω-datapoints on $U^*(q_\alpha)$ is*

$$q_\alpha(\omega) \left(\sum_{\omega' \in \Omega} v_{q_\alpha}^*(\omega') \sigma_\omega(\omega') - v_{q_\alpha}^*(\omega) \right).$$

This effect is decreasing in α. Finally, such refinements always weakly benefit the principal: $U^*(q_\alpha) \geq U^*(q)$ for all α.

The first part of the result shows that the overall positive effects on the refined datapoints dominates also for the overall payoff of the principal. Again, this is not immediate since changes in q_α can reduce the value of some types of datapoints, whether they are affected by the refinement or not. The second part of the result shows that at the margin, as the principal learns more about its database by refining ω-datapoints, she benefits less. This is because the expected value gain of the marginal datapoint becomes smaller. This part is akin to a decreasing marginal value of information, but with some qualifications. Note that here the principal is not getting marginally more information about one single interaction, which may correspond to the usual way of thinking about the marginal value of information in a single decision problem. Rather, she is getting the same information about a marginal interaction and hence about the whole database, which changes her informational advantage for the overall mediation problem.

One may wonder whether these results about the value of information for mediation problems depend on how datapoints are refined. We can describe general refinements as follows. For every q and $\omega \in \Omega$, let $Q(\omega, q)$ be the set of all databases that can be reached from q by refining ω-datapoints: $Q(\omega, q)$ contains all $q' \in \mathbb{R}_+^\Omega$ that satisfy

- $q'(\omega) \leq q(\omega)$,
- $q'(\hat{\omega}) \geq q(\hat{\omega})$ if $\hat{\omega} \subset \omega$,
- $q'(\hat{\omega}) = q(\hat{\omega})$ if $\hat{\omega} \not\subseteq \omega$,
- $\sum_{\hat{\omega} \in \Omega} q'(\hat{\omega}) = \sum_{\hat{\omega} \in \Omega} q(\hat{\omega})$.

Then, given q, acquiring better data about ω-datapoints involves transitioning—possibly with some randomness—from q to some $q' \in Q(\omega, q)$. That is, any such data-acquisition process can be described by some distribution $\rho(\omega, q) \in \Delta(Q(\omega, q))$. While i.i.d. refinements induce such a distribution, other refinements may involve correlation in how the principal learns about different ω-datapoints. It is easy to see that for full-disclosure problems *any* refinement $\rho(\omega, q)$ increases, in expectation, both the value of the refined datapoints and the principal's total payoff.[22] This is analogous to the result that the value of information in decision problems is always non-negative. However, this can fail for other mediation problems, as the next example illustrates. The basic reason is that our acquisition of information changes not only how much the principal knows about each interaction, but also the degree and nature of the asymmetric information between the principal and the agents implied by the commonly-known database composition.[23] This highlights another important distinctive feature of the value of data in mediation problems.

Example. (Cont'd) Suppose that $h > \frac{1}{2}$ and the initial q satisfies condition (ii) in Table 1. In addition, assume that $q(1) + q(\omega^\circ)(1-h) > q(2)$. Consider the following general refinement $\rho(\omega^\circ, q)$, which is arguably extreme but serves to make our point as clearly as possible. Suppose the platform is told that all its ω°-datapoints involve buyers with the same valuation. Thus, if refined, with probability $1-h$ they *all* become 1-datapoints and with probability h they *all* become 2-datapoints. That is, $\rho(\omega^\circ, q)$ assigns positive probability to only q' and q'', where $q'(1) = q(1) + q(\omega^\circ)$, $q'(2) = q(2)$, $q''(1) = q(1)$, $q''(2) = q(2) + q(\omega^\circ)$, and $q'(\omega^\circ) = q''(\omega^\circ) = 0$. From Table 1, we have $v^*_{q'}(1) = 0$ and $v^*_{q''}(2) = 0$, which implies that

$$(1-h)v^*_{q'}(1) + hv^*_{q''}(2) - v^*_q(\omega^\circ) = -(1-h) < 0.$$

It is also easy to calculate that

$$U^*(q) = q(1) + q(\omega^\circ)(1-h) > \max\{q(2), q(1)\} = \max\{U^*(q'), U^*(q'')\}.$$

Therefore, this refinement has a strictly negative effect not only on the value of the refined ω°-datapoints, but also on the platform's overall payoff. Note that if the platform maximizes the seller's profits, $v^*_{\hat{q}}(1) = 1$, $v^*_{\hat{q}}(2) = 2$, and $v^*_{\hat{q}}(\omega^\circ) = 2h$ for all $\hat{q}$. Therefore, the *same* refinement $\rho(\omega^\circ, q)$ has a strictly positive effect on both the value of ω°-datapoints and the platform's overall payoff, as standard in decision problems.

The key is that a profit-maximizing platform treats each buyer-seller interaction independently and hence does not care about correlation in how it learns about datapoints. By contrast,

[22] This follows from the fact that $v^*_q(\omega)$ is independent of q for all $\omega \in \Omega$ and therefore only the marginal of ρ for each refined datapoint matters.

[23] Of course, in reality the principal may acquire information privately without the agents' knowledge. This introduces further complications, which we leave for future research.

a profit-maximizing platform cares about such correlation, because it can have profound consequences on its informational advantage through the composition of its database. $\triangle$

To recap, if given the option to acquire better data about ω-datapoints by drawing at random more precise observations about each datapoint, the principal always has a non-negative willingness to pay for such information, which may be decreasing in the scope of learning. However, she may strictly prefer to not acquire better data that involves high correlation across ω-datapoints. In practice, for large databases it may be more reasonable that acquiring better data takes the form of random draws from a large population. The conceptual point remains that, unlike for decision-makers, for mediators information can have negative value.

5 A Gambling Perspective on Data Value and Externalities

To better understand the value of datapoints and the externalities between them, it will help to provide a stand-alone interpretation of the data-value problem $\mathcal{V}_q$ explaining what (b, ℓ) are and how the principal uses them. With minor adjustments, this interpretation applies to problems where the agents observe some data and the principal takes some action. In this part, we will fix $q \in \mathbb{R}^{\Omega}_{++}$ and so drop it from notation.

5.1 Gambles Against the Agents

Our interpretation hinges on unpacking how the principal determines the agents' contributions to the externalities between datapoints. The value formula (3) reveals that she does so through her choice of b and ℓ, which fully pin down each $t(a, \omega)$ and hence ultimately $v(\omega)$. Recall that the principal wants to minimize the values of datapoints, so she would like to lower $t(a, \omega) = \sum_{i \in I} t_i(a, \omega)$ as much as possible for all (a, ω). Each term of $t_i(a, \omega)$ takes the form

$$b_i(a_i)\ell_i(a_i'|a_i)\left(u_i(a_i, a_{-i}, \omega) - u_i(a_i', a_{-i}, \omega)\right),$$

which contributes to lowering $t_i(a, \omega)$ if and only if $\ell_i(a_i'|a_i) > 0$ and $u_i(a_i, a_{-i}, \omega) < u_i(a_i', a_{-i}, \omega)$. In words, if agent i knew his interaction's type ω and his opponents' actions a_{-i}, he would strictly prefer a_i' to a_i. In this case, playing a_i amounts to making a mistake from an ex-post viewpoint. We will then say that agent i regrets playing a_i.

Thus, inducing agents to play actions they will regret emerges as an intrinsic goal of the principal's problem—together with maximizing her payoff u_0 of course. In this view, (b_i, ℓ_i) and the corresponding t_i become an exploitation strategy on the part of the principal against agent i. In order to induce regrettable actions, she must withhold some information from agent i about ω or a_{-i}. This explains why the principal may prefer partial disclosure, from the perspective of her data-value problem. In the end, the value $v(\omega)$ results from a trade-off between the payoff $u_0(a, \omega)$ and the return from inducing agents to choose regrettable actions.

This return depends on the structure of b and ℓ, which define a family of gambles against the agents. To see this, fix any (a, ω) and agent i. Then, $\ell_i(\cdot|a_i) \in \Delta(A_i)$ defines a lottery over prizes, where for each a_i' the prize is $u_i(a_i, a_{-i}, \omega) - u_i(a_i', a_{-i}, \omega)$ for the principal; the scaling term $b_i(a_i)$ defines the stakes that she bets on this lottery. Given her objective, the principal "wins" when $u_i(a_i, a_{-i}, \omega) < u_i(a_i', a_{-i}, \omega)$ and "loses" otherwise. Thus, we can interpret $t(a, \omega)$ as the overall expected prize from (b, ℓ). We can then think of $\mathcal{V}$ as a fictitious environment where money is a medium of exchange and the principal can write monetary gambling contracts with each agent. Such contracts are enforced through contingent-claim markets that determine prizes based on the interaction's type ω and outcome a.[24]

We can then link the externalities between datapoints with how the principal chooses these gambles in $\mathcal{V}$. Negative externalities $t^* < 0$ correspond to gambles favorable to the principal, in the sense that wins exceed losses in expectation. This requires the help of other datapoints to withhold information and thus induce the agents to play actions they will regret. Conversely, positive externalities $t^* > 0$ correspond to gambles unfavorable to the principal, in the sense that losses exceed wins in expectation. Corollary 1 implies that, at the optimum, the principal chooses gambles that favor her for some datapoints, but not for others. In fact, this stems from deeper constraints and trade-offs in the use of such gambles against the agents.

5.2 Feasible Gambles and Trade-offs

The gambles the principal can use to exploit the agents in $\mathcal{V}$ have specific features that help us understand the nature of the data-value problem.

Some of these features reflect structural properties of $\mathcal{V}$. While the prizes of each gamble are contingent on both ω and the entire a, for each agent i both b_i and ℓ_i can depend only of his a_i. This constrains the principal's ability to tailor her gambles with each agent across datapoints. These properties reflect in $\mathcal{V}$ key interdependences in $\mathcal{U}$: The independence of (b_i, ℓ_i) from a_{-i} reflects the interdependence in $\mathcal{U}$ between agents' incentives; the independence of (b_i, ℓ_i) from ω reflects the non-separability of $\mathcal{U}$ across datapoints. To see this, suppose $\ell_i(\hat{a}_i|a_i) > 0$. Then, (b_i, ℓ_i) links the right-hand side of the value formula (3) for (a_i, a_{-i}, ω) and (a_i, a'_{-i}, ω').

[24] See Nau (1992) for a related interpretation.

In particular, if $u_i(a_i, a_{-i}, \omega) < u_i(\hat{a}_i, a_{-i}, \omega)$ but $u_i(a_i, a'_{-i}, \omega') > u_i(\hat{a}_i, a'_{-i}, \omega')$, the principal faces a trade-off in determining v, as she may not be able to use (b_i, ℓ_i) to lower $v(\omega)$ without also raising $v(\omega')$. This is another way to see why and how externalities arise between datapoints. When committing to (b, ℓ) the principal has to take into account these effects of each (b_i, ℓ_i) across datapoints. How she solves these trade-offs depends on the relative frequency of datapoints in the database (hence q). Importantly, this transformation of non-separabilities in $\mathcal{U}$ into independence properties of (b, ℓ) is what enables $\mathcal{V}$ to assign values individually to each datapoint.

In fact, the principal faces other restrictions in her ability to *jointly* exploit the agents. It is intuitive that she would want to design (b, ℓ) so that $t(a, \omega) \leq 0$ for all (a, ω) with some strict inequality. This would guarantee a sure arbitrage against the agents. Such gambles, however, are infeasible in the following sense. Recall that by complementary slackness $x^*(a|\omega) > 0$ implies $v^*(\omega) = u_0(a, \omega) + t^*(a, \omega)$. Thus, since every ω must induce some action profile for every x, action profiles that cannot be in the support of any obedient $x(\cdot|\omega)$ are irrelevant for determining $v^*(\omega)$. Given this, define

$$\mathbf{X} = \{(a, \omega) \in A \times \Omega : x(a|\omega) > 0 \text{ for some obedient } x\}.$$

Let $G(\mathbf{X})$ be the set of gambles that can be contingent only on pairs $(a, \omega) \in \mathbf{X}$ (formally, we restrict the functions b and ℓ to the subdomain $\mathbf{X}$). Note that restricting the principal to choosing from $G(\mathbf{X})$ in $\mathcal{V}$ is immaterial for its optimal solution, in the same way that restricting x to the domain $\mathbf{X}$ is immaterial in $\mathcal{U}$.

Proposition 7. *For every gamble $(b, \ell) \in G(\mathbf{X})$, if $t(a, \omega) < 0$ for some (a, ω), there must exist (a', ω') such that $t(a', \omega') > 0$.*

This property is closely related to a very similar result in Nau (1992). For completeness Appendix B provides a proof, which relies on a dual characterization of $\mathbf{X}$ using Farkas' lemma.

The economic takeaway is that in her attempt to minimize values v by exploiting the agents with (b, ℓ), the principal faces a fundamental trade-off, which is a hallmark of problem $\mathcal{V}$. Successfully exploiting the agents for some type ω of datapoints with some outcome a requires paying the cost of losing against them for some other type ω' of datapoints or outcome a'. Note that this result is stronger than Corollary 1, as it refers to the deep structure of data-value problems for mediators. It also sheds light on how and how much they can actually manipulate agents by conveying information.

References

Acemoglu, D., A. Makhdoumi, A. Malekian, and A. Ozdaglar (2021): "Too Much Data: Prices and Inefficiencies in Data Markets," *Americal Economic Journal: Microeconomics*, Forthcoming.

Acquisti, A., C. Taylor, and L. Wagman (2016): "The Economics of Privacy," *Journal of Economic Literature*, 54, 442–92.

Admati, A. and P. Pfleiderer (1990): "Direct and Indirect Sale of Information," *Econometrica*, 58, 901–28.

Admati, A. R. and P. Pfleiderer (1986): "A Monopolistic Market for Information," *Journal of Economic Theory*, 39, 400–438.

Ali, S. N., G. Lewis, and S. Vasserman (2020): "Voluntary Disclosure and Personalized Pricing," *arXiv preprint arXiv:1912.04774v2*.

Arrieta-Ibarra, I., L. Goff, D. Jiménez-Hernández, J. Lanier, and E. G. Weyl (2018): "Should We Treat Data as Labor? Moving beyond "Free"," *AEA Papers and Proceedings*, 108, 38–42.

Bergemann, D. and A. Bonatti (2015): "Selling Cookies," *American Economic Journal: Microeconomics*, 7, 259–94.

——— (2019): "Markets for Information: An Introduction," *Annual Review of Economics*, 11, 85–107.

Bergemann, D., A. Bonatti, and T. Gan (2021): "The Economics of Social Data," *arXiv preprint arXiv:2004.03107v2*.

Bergemann, D., A. Bonatti, and A. Smolin (2018): "The Design and Price of Information," *American Economic Review*, 108, 1–48.

Bergemann, D., B. Brooks, and S. Morris (2015): "The Limits of Price Discrimination," *American Economic Review*, 105 (3).

Bergemann, D. and S. Morris (2016): "Bayes Correlated Equilibrium and the Comparison of Information Structures in Games," *Theoretical Economics*, 11, 487–522.

——— (2019): "Information Design: A Unified Perspective," *Journal of Economic Literature*, 57(1), pp. 44-95).

Bertsimas, D. and J. N. Tsitsiklis (1997): *Introduction to Linear Optimization*, Athena Scientific.

Brooks, B. and S. Du (2020): "A Strong Minimax Theorem for Informationally-Robust Auction Design," *Working Paper*.

——— (2021): "Optimal Auction Design with Common Values: An Informationally-Robust Approach," *Econometrica*, 89(3), 1313–1360.

Budzyn, A. (2019): "Data is the Oil of the Digital World. What if Tech Giants Had to Buy it from Us?" *World Economic Forum*, https://www.weforum.org/agenda/2019/04/data-oil-digital-world-asset-tech-giants-buy-it/.

Calzolari, G. and A. Pavan (2006): "On the Optimality of Privacy in Sequential Contracting," *Journal of Economic Theory*, 130, 168–204.

Das, S., E. Kamenica, and R. Mirka (2017): "Reducing Congestion through Information Design," in *2017 55th annual allerton conference on communication, control, and computing (allerton)*, IEEE, 1279–1284.

Dizdar, D. and E. Kováč (2020): "A Simple Proof of Strong Duality in the Linear Persuasion Problem," *Games and Economic Behavior*, 122, 407–412.

Dorfman, R., P. A. Samuelson, and R. M. Solow (1987): *Linear Programming and Economic Analysis*, Courier Corporation.

Du, S. (2018): "Robust Mechanisms Under Common Valuation," *Econometrica*, 86(5), 1569–1588.

Dworczak, P. and A. Kolotilin (2019): "The Persuasion Duality," *Available at SSRN 3474376*.

Dworczak, P. and G. Martini (2019): "The Simple Economics of Optimal Persuasion," *Journal of Political Economy*, 127, 1993–2048.

Elliott, M., A. Galeotti, and A. Koh (2020): "Market Segmentation through Information," *Working Paper*.

Gale, D. (1989): *The Theory of Linear Economic Models*, University of Chicago press.

Galperti, S. and J. Perego (2018): "A Dual Perspective on Information Design," *Available at SSRN 3297406*.

Gentzkow, M. and E. Kamenica (2016): "Competition in Persuasion," *Review of Economic Studies*, 84, 300–322.

Kamenica, E. and M. Gentzkow (2011): "Bayesian Persuasion," *American Economic Review*, 101, 2590 – 2615.

Kolotilin, A. (2018): "Optimal Information Disclosure: A Linear Programming Approach," *Theoretical Economics*, 13, 607 – 635.

Lanier, J. (2013): *Who Owns the Future?*, Simon & Schuster.

Maskin, E. and J. Tirole (1990): "The Principal-Agent Relationship with an Informed Principal: The Case of Private Values," *Econometrica*, 58, 379–409.

Mathevet, L., J. Perego, and I. Taneva (2020): "On Information Design in Games," *Journal of Political Economy*, 128, 1370–1404.

Milgrom, P. and C. Shannon (1994): "Monotone Comparative Statics," *Econometrica*, 157–180.

Myerson, R. B. (1983): "Mechanism Design by an Informed Principal," *Econometrica*, 51, 1767–1797.

——— (1984): "Two-Person Bargaining Problems with Incomplete Information," *Econometrica*, 52, 461–488.

——— (1997): "Dual Reduction and Elementary Games," *Games and Economic Behavior*, 21, 183–202.

Nau, R. F. (1992): "Joint Coherence in Games of Incomplete Information," *Management Science*, 38, 374–387.

Nau, R. F. and K. F. McCardle (1990): "Coherent Behavior in Noncooperative Games," *Journal of Economic Theory*, 50, 424–444.

Posner, E. and E. G. Weyl (2018): *Radical Markets: Uprooting Capitalism and Democracy for a Just Society*, Princeton University Press.

Rockafellar, R. T. (1970): *Convex analysis*, vol. 36, Princeton university press.

Appendix

A A Sufficient Condition for Suboptimality of Full Disclosure

We provide a sufficient condition on Γ for suboptimality of full disclosure for the general case where the principal can choose $a_0 \in A_0$ and each agent i can privately observe some own data $\omega_i \in \Omega_i$ about the interaction he is in. Recall that if the principal fully reveals all ω, then she must be implementing a correlated equilibrium of the complete-information game Γ_ω for all ω, i.e., $x_q^*(\cdot|\omega) \in CE(\Gamma_\omega)$. The definition of CE in terms of inequalities can be adjusted to incorporate the principal's a_0.

Proposition 8. *Fix Γ. Suppose there exists (a, ω) that satisfies:*

(1) $u_0(a, \omega) > \bar{u}(\omega)$,

(2) for every agent i and action $\hat{a}_i$, such that $u_i(a_i, a_{-i}, \omega) < u_i(\hat{a}_i, a_{-i}, \omega)$, there exists an $x(\cdot|\omega') \in CE(\Gamma_{\omega'})$ for some ω', with $\omega'_i = \omega_i$, that satisfies

$$\sum_{a \in A} u_0(a, \omega') x(a|\omega') = \bar{u}(\omega'),$$

$$\sum_{a_{-i} \in A_{-i}} \left(u_i(a_i, a_{-i}, \omega') - u_i(\hat{a}_i, a_{-i}, \omega')\right) x(a_i, a_{-i}|\omega') > 0.$$

Then $\mathcal{U}_q$ does not admit an FIC solution for any $q \in \mathbb{R}^{\Omega}_{++}$.

Condition (1) is clearly necessary: If for every datapoint ω every action profile a cannot deliver a payoff higher than the full-information payoff $\bar{u}(\omega)$, then it is clearly optimal for the principal to fully reveal every ω. Given an outcome (a, ω) with $u_0(a, \omega) > \bar{u}(\omega)$, there must be an agent who would have a profitable deviation from a_i to $\hat{a}_i$ if he knew (a_{-i}, ω_{-i}). Otherwise, given a_0, the profile a_{-0} is a Nash Equilibrium of Γ_ω and hence $a_{-0} \in CE(\Gamma_\omega)$, which would imply $u_0(a, \omega) \leq \bar{u}(\omega)$. Then condition (2) requires that agent i's data ω_i is consistent with another datapoint ω'—so that he cannot tell ω and ω' apart based on his own data only—which admits a principal-preferred correlated equilibrium that also recommends i to play a_i and renders the deviation to $\hat{a}_i$ strictly suboptimal.

B Proofs

Proof of Lemma 1. We will prove the result for the general case where the agents observe private data in the form of $\omega_i \in \Omega_i$ and hence $\omega = (\omega_0, \omega_1, \dots, \omega_n)$ (see Section 2.2). The special case where only the principal observes data obtains by having $|\Omega_i| = 1$ for all $i \in I$. We will also forumlate the problem directly in terms of choosing a measure $\chi \in \mathbb{R}_+^{A\times\Omega}$. Formally, the problem is

$$\begin{aligned}
\mathcal{U}_q: \qquad & \max_{\chi} \sum_{\omega\in\Omega, a\in A} u_0(a,\omega)\chi(a,\omega) \\
& \text{s.t. for all } i \in I, \omega_i \in \Omega_i, \text{ and } a_i, a_i' \in A_i, \\
& \sum_{\omega_{-i}\in\Omega_{-i}, a_{-i}\in A_{-i}} \left(u_i(a_i,a_{-i},\omega) - u_i(a_i',a_{-i},\omega)\right)\chi(a_i,a_{-i},\omega) \geq 0, \qquad (8) \\
& \text{and for all } \omega \in \Omega, \qquad (9) \\
& \quad \sum_{a\in A} \chi(a,\omega) = q(\omega).
\end{aligned}$$

It is convenient to express this problem in matrix form. Fix an arbitrary total ordering of the set $A \times \Omega$. We denote by $\mathbf{u}_0 \in \mathbb{R}^{A\times\Omega}$ the vector whose entry corresponding to (a,ω) is $u_0(a,\omega)$. For every player i, let $\mathbf{U}_i \in \mathbb{R}^{(A_i\times A_i\times\Omega_i)\times(A\times\Omega)}$ be a matrix thus defined: For each row $(a_i', a_i'', \omega_i') \in A_i \times A_i \times \Omega_i$ and column $(a,\omega) \in A \times \Omega$, let the corresponding entry be

$$\mathbf{U}_i((a_i',a_i'',\omega_i'),(a,\omega)) = \begin{cases} u_i(a_i',a_{-i},\omega) - u_i(a_i'',a_{-i},\omega) & \text{if } a_i' = a_i, \omega_i' = \omega_i \\ 0 & \text{else.} \end{cases}$$

Thus, $\mathbf{U}_i(a_i',a_i'',\omega_i')$ denotes the row labeled by (a_i',a_i'',ω_i') (which defines the corresponding obedience constraint) and $\mathbf{U}_i(a,\omega)$ denotes the column labeled by (a,ω). Define the matrix $\mathbf{U}$ by stacking all the matrices $\{\mathbf{U}_i\}_{i\in I}$ on top each other. Finally, define the indicator matrix $I \in \{0,1\}^{\Omega\times(A\times\Omega)}$ such that, for each row ω' and column (a,ω'),

$$I(\omega',(a,\omega)) := \begin{cases} 1 & \text{if } \omega' = \omega \\ 0 & \text{else.} \end{cases}$$

With this notation and treating q as a vector, $\mathcal{U}_q$ can be written as follows:

$$\begin{aligned}
\max_{\chi} \quad & \mathbf{u}_0^T\chi \\
\text{s.t.} \quad & \mathbf{U}\chi \geq \mathbf{0}, \\
& I\chi = q, \qquad (10) \\
& \chi \geq \mathbf{0}.
\end{aligned}$$

Given this, by standard linear-programming arguments (Bertsimas and Tsitsiklis (1997)) the dual of $\mathcal{U}_q$ can be written as

$$\min_{\lambda,v} \mathbf{0}^T \lambda + q^T v$$

subject to for all $i = 1, \ldots, n$, $a_i, a_i' \in A_i$, and $\omega_i \in \Omega_i$,

$$\lambda_i(a_i'|a_i, \omega_i) \geq 0,$$

$v(\omega) \in \mathbb{R}$ for all $\omega \in \Omega$ (i.e., it is unconstrained), and for all $(a, \omega) \in A \times \Omega$

$$u_0(a, \omega) \leq v(\omega) - \sum_{i \in I} \left\{ \sum_{a_i' \in A_i} \left(u_i(a_i, a_{-i}, \omega) - u_i(a_i', a_{-i}, \omega) \right) \lambda_i(a_i'|a_i, \omega_i) \right\}.$$

The objective simplifies to

$$\min_{\lambda,v} \sum_{\omega \in \Omega} v(\omega) q(\omega).$$

The second set of constraints can be written as

$$v(\omega) \geq u_0(a, \omega) + \sum_{i \in I} \left\{ \sum_{a_i' \in A_i} \left(u_i(a_i, a_{-i}, \omega) - u_i(a_i', a_{-i}, \omega) \right) \lambda_i(a_i'|a_i, \omega_i) \right\}.$$

Finally, we express this dual in a form that is equivalent to $\mathcal{V}_q$ by exploiting the structure of the specific problem at hand. To this end, for every i and ω_i we can set $\lambda_i(a_i|a_i, \omega_i) = 1$ (or any strictly positive value) for all $a_i \in A_i$. Given this, for every i and $(a_i, \omega_i) \in A_i \times \Omega_i$, define

$$b_i(a_i, \omega_i) = \sum_{a_i' \in A_i} \lambda_i(a_i'|a_i, \omega_i),$$

which is strictly positive by construction. Also, for every i and $(a_i', a_i, \omega_i) \in A_i \times A_i \times \Omega_i$ define

$$\ell_i(a_i'|a_i, \omega_i) = \frac{\lambda_i(a_i'|a_i, \omega_i)}{b_i(a_i, \omega_i)},$$

which implies that $\ell_i(\cdot|a_i, \omega_i) \in \Delta(A_i)$. Letting $b = (b_1, \ldots, b_n)$ and $\ell = (\ell_1, \ldots, \ell_n)$ so defined, we have the dual of $\mathcal{U}_q$ is equivalent to

$$\min_{v,b,\ell} \sum_{\omega \in \Omega} v(\omega) q(\omega)$$

subject to for all (a, ω)

$$v(\omega) \geq u_0(a, \omega) + \sum_{i \in I} t_i(a, \omega),$$

where

$$t_i(a,\omega) = b_i(a_i,\omega_i) \sum_{a_i' \in A_i} \left(u_i(a_i,a_{-i},\omega) - u_i(a_i',a_{-i},\omega)\right) \ell_i(a_i'|a_i,\omega_i).$$

Since for every $\omega \in \Omega$ this constraint has to hold for all $a \in A$ and the data-value problem is the minimization problem, we conclude that each $v(\omega)$ has to satisfy

$$v(\omega) = \max_{a \in A} \{u_0(a,\omega) + t(a,\omega)\},$$

where $t(a,\omega) = \sum_{i \in I} t_i(a,\omega)$. Thus, we obtain problem $\mathcal{V}_q$. □

Remark 1. *We can transform $\mathcal{U}_q$ to the standard form $\mathcal{U}_q^S$ which can be written as follows:*

$$\begin{array}{rrcl} \max_{\chi,s} & \mathbf{u}_0\chi & & \\ s.t. & \mathbf{U}\chi - s & = & \mathbf{0}, \\ & I\chi & = & q, \\ & \chi, s & \geq & \mathbf{0}, \end{array} \tag{11}$$

where each $s_i(a_i'|a_i,\omega_i)$ is a nonnegative slack variable. The dual of $\mathcal{U}_q^S$ coincides with the data-value problem $\mathcal{V}_q$. Note that $\mathcal{U}_q$ always has an optimal solution χ_q^*, which is generically unique and hence corresponds to an extreme point of the polyhedron of feasible χ. Moreover, this χ_q^* is an optimal solution of $\mathcal{U}_q^S$ as well. The extreme point χ_q^* is nondegenerate by Assumption 1 and characterized by a square nonsingular active-constraint submatrix $\mathbf{B}$ consisting of linearly independent rows of the stacked matrix $\left[\begin{array}{c|c} \mathbf{U} & -\mathbf{1} \\ \hline I & \mathbf{0} \end{array}\right]$, where $\mathbf{1}$ is the identity matrix. As illustrated in Chapter 4 of Bertsimas and Tsitsiklis (1997), given $\mathbf{B}$, we have

$$\left[\begin{array}{c} \chi_q^* \\ \hline s_q^* \end{array}\right] = \mathbf{B}^{-1} \left[\begin{array}{c} \mathbf{0} \\ \hline q \end{array}\right], \tag{12}$$

where s_q^* is the vector of optimal slack variables in $\mathcal{U}_q^S$. A corresponding solution of $\mathcal{V}_q$ is given by

$$\left[\begin{array}{c} v_q^* \\ \hline \lambda_q^* \end{array}\right] = \mathbf{u}_0\mathbf{B}^{-1}. \tag{13}$$

It follows that as long as the optimal solutions of $\mathcal{U}_q$ and $\mathcal{V}_q$ are defined by the same extreme point defined by $\mathbf{B}$, χ_q^* varies with q, but (v_q^*, λ_q^*) does not.

Proof of Lemma 2. Fix an optimal solution (v_q^*, b_q^*, ℓ_q^*) of $\mathcal{V}_q$. For every q, $\omega \in \Omega$, and $x(\cdot|\omega) \in CE(\Gamma_\omega)$, by (3) we have

$$v_q^*(\omega) \geq \sum_{a \in A} u_0(a,\omega)x(a|\omega) + \sum_{a \in A} t(a,\omega)x(a|\omega)$$

$$
\begin{aligned}
&= \sum_{a\in A} u_0(a,\omega)x(a|\omega) \\
&\quad + \sum_{a\in A}\left\{\sum_{i\in I} b_i^*(a_i,\omega_i)\sum_{\hat{a}_i\in A_i}\left(u_i(a_i,a_{-i},\omega)-u_i(\hat{a}_i,a_{-i},\omega)\right)\ell_i^*(\hat{a}_i|a_i,\omega_i)\right\}x(a|\omega) \\
&= \sum_{a\in A} u_0(a,\omega)x(a|\omega) \\
&\quad + \sum_{i\in I}\sum_{a_i,\hat{a}_i\in A_i} b_i^*(a_i,\omega_i)\ell_i^*(\hat{a}_i|a_i,\omega_i)\left\{\sum_{a_{-i}\in A_{-i}}\left(u_i(a_i,a_{-i},\omega)-u_i(\hat{a}_i,a_{-i},\omega)\right)x(a|\omega)\right\} \\
&\geq \sum_{a\in A} u_0(a,\omega)x(a|\omega),
\end{aligned}
$$

where the last inequality follows because any $x(\cdot|\omega) \in CE(\Gamma_\omega)$ is defined by the property that, for all $i \in I$ and $a_i, a_i' \in A_i$,

$$
\sum_{a_{-i}\in A_{-i}}\left(u_i(a_i,a_{-i},\omega)-u_i(a_i',a_{-i},\omega)\right)x(a_i,a_{-i}|\omega) \geq 0.
$$

Since $x(\cdot|\omega)$ is an arbitrary element of $CE(\Gamma_\omega)$, we conclude that $v_q^*(\omega) \geq \bar{u}(\omega)$. □

Proof of Proposition 1. By complementary slackness, $x_q^*(a,\omega) > 0$ implies $v_q^*(\omega) = u_0(a,\omega) + t_q^*(a,\omega)$. Hence,

$$
v_q^*(\omega) = \sum_{a\in A} u_0(a,\omega)x_q^*(a|\omega) + \sum_{a\in A} t_q^*(a,\omega)x_q^*(a|\omega) = u_q^*(\omega) + t_q^*(\omega).
$$

Suppose we start from database q, with $q(\omega) > 0$, and we increase the quantity of ω-datapoints from $q(\omega)$ to $\hat{q}(\omega)$, thus obtaining the database $\hat{q}$. We can write

$$
U^*(\hat{q}) - U^*(q) = u_{\hat{q}}^*(\omega)[\hat{q}(\omega)-q(\omega)] + \sum_{\omega'\in\Omega}[u_{\hat{q}}^*(\omega') - u_q^*(\omega')]\hat{q}(\omega')
$$

Dividing both sides by $\hat{q}(\omega) - q(\omega)$, taking limits as $\hat{q}(\omega) \to q(\omega)$, and using Lemma 1, we obtain that

$$
\begin{aligned}
t_q^*(\omega) &= v_q^*(\omega) - u_q^*(\omega) = \frac{\partial U^*(q)}{\partial q(\omega)} - u_q^*(\omega) \\
&= \lim_{\hat{q}(\omega)\to q(\omega)} \frac{\sum_{\omega'\in\Omega}[u_{\hat{q}}^*(\omega') - u_q^*(\omega')]\hat{q}(\omega')}{\hat{q}(\omega)-q(\omega)} = \sum_{\omega'\in\Omega}\frac{\partial u_q^*(\omega')}{\partial q(\omega)}q(\omega') \\
&= \sum_{\omega'\in\Omega, a\in A} u_0(a,\omega')\left(\lim_{\hat{q}(\omega)\to q(\omega)}\frac{[x_{\hat{q}}^*(a|\omega') - x_q^*(a|\omega')]}{\hat{q}(\omega)-q(\omega)}\right)\hat{q}(\omega') = \\
&= \sum_{\omega'\in\Omega, a\in A} u_0(a,\omega')\frac{\partial x_q^*(a|\omega')}{\partial q(\omega)}q(\omega'),
\end{aligned}
$$

where the existence of the derivative $\frac{\partial x_q^*(a|\omega')}{\partial q(\omega)}$ almost everywhere follows from (12). □

Proof Proposition 2. By the formulation of $\mathcal{V}_q$ and Lemma 2, the polyhedron of feasible solutions of $\mathcal{V}_q$, denoted by $F(\mathcal{V}_q)$ does not contain a line because all dual variables are bounded from below. By Theorem 2.6 in Bertsimas and Tsitsiklis (1997), $F(\mathcal{V}_q)$ has at least one extreme point and at most finitely many of them by Corollary 2.1 in Bertsimas and Tsitsiklis (1997). By Theorem 4.4 in Bertsimas and Tsitsiklis (1997), $\mathcal{V}_q$ has at least one optimal solution. By Theorem 2.7 in Bertsimas and Tsitsiklis (1997), we can focus on solutions that are extreme points of $F(\mathcal{V}_q)$.

Fix q and suppose that the optimal solution (v_q^*, λ_q^*) of the dual of $\mathcal{U}_q$ is unique. As explained in Remark 1, there exists a submatrix $\mathbf{B}$ such that (v_q^*, λ_q^*) satisfies (13). Given Assumption 1, Theorem 3.1 and Exercise 3.6 in Bertsimas and Tsitsiklis (1997) imply that

$$\left[\begin{array}{c|c} \mathbf{U} & -\mathbf{1} \\ \hline I & \mathbf{0} \end{array}\right] \mathbf{B}^{-1} \left[\begin{array}{c} \mathbf{0} \\ \hline q \end{array}\right] \geq \left[\begin{array}{c} \mathbf{0} \\ \hline q \end{array}\right].$$

The inequality is strict for each row of $\mathbf{U}$ that corresponds to $\lambda_{q,i}^*(a_i'|a_i, \omega_i) = 0$ (or, equivalently, $\ell_{q,i}^*(a_i'|a_i, \omega_i) = 0$):

$$\left[\mathbf{U}_i(a_i, a_i', \omega_i) \mid -\mathbf{1}_i(a_i, a_i', \omega_i)\right] \mathbf{B}^{-1} \left[\begin{array}{c} \mathbf{0} \\ \hline q \end{array}\right] > 0, \tag{14}$$

where $\mathbf{1}_i(a_i, a_i', \omega_i)$ is the row of the identity matrix $\mathbf{1}$ that corresponds to (i, a_i, a_i', ω_i). Note that for each row ω of the indicator matrix I (i.e., $I(\omega)$), which corresponds to variable $v_q^*(\omega)$, it automatically holds that $\left[I(\omega) \mid \mathbf{0}\right] \mathbf{B}^{-1} \left[\begin{smallmatrix} \mathbf{0} \\ q \end{smallmatrix}\right] = q(\omega)$. Similarly, for each row of $\mathbf{U}$ that corresponds to $\lambda_{q,i}^*(a_i'|a_i, \omega_i) > 0$ (or, equivalently, $\ell_{q,i}^*(a_i'|a_i, \omega_i) > 0$), it holds that $\left[\mathbf{U}_i(a_i, a_i', \omega_i) \mid -\mathbf{1}_i(a_i, a_i', \omega_i)\right] \mathbf{B}^{-1} \left[\begin{smallmatrix} \mathbf{0} \\ q \end{smallmatrix}\right] = 0$ as long as $\mathbf{B}$ identifies the optimal extreme point.

Now consider changes in q and note that it only enters the objective of $\mathcal{V}_q$. Each condition (14) defines an open set of q's in $\mathbb{R}_+^\Omega$ that satisfy it. Define $(v_{\mathbf{B}}^*, \lambda_{\mathbf{B}}^*)$ identified by $\mathbf{B}$ as in (13) and

$$Q(\mathbf{B}) = \{q : (14) \text{ holds for all } i \in I \text{ and } (a_i, a_i', \omega_i) \text{ s.t. } \lambda_{\mathbf{B},i}^*(a_i|a_i', \omega_i) = 0\}.$$

Note that $Q(\mathbf{B})$ is an open set because it is the intersection of finitely many open sets.

Now recall that there are only finitely many extreme points of the dual polyhedron of feasible solutions. Therefore, there are finitely many submatrices $\{\mathbf{B}_1, \ldots, \mathbf{B}_K\}$ such that each identifies an optimal $(v_{\mathbf{B}_k}^*, \lambda_{\mathbf{B}_k}^*)$ that is unique for all $q \in Q(\mathbf{B}_k)$. For all $k = 1, \ldots, K$, define $Q_k = Q(\mathbf{B}_k)$. By construction, each Q_k is open and $q, q' \in Q_k$ implies that $(v_q^*, \lambda_q^*) = (v_{q'}^*, \lambda_{q'}^*)$.

Since (v_q^*, λ_q^*) is generically unique with respect to q, it follows that $\mathbb{R}_+^\Omega \setminus \cup_k Q_k$ has Lebesgue measure zero. □

Proof of Proposition 3. Fix $\mu_1, \mu_2 \in \Delta(\Omega)$. Let $\Omega^i = \{\omega \in \Omega : \mu_i(\omega) > \mu_j(\omega), j \neq i\}$, $i \in \{1,2\}$, and $\Omega^3 = \Omega \setminus \Omega_1 \setminus \Omega_2$.

Let $X = \mathbb{R}^\Omega \times \mathbb{R}_+^{A_1 \times A_1} \times \ldots \times \mathbb{R}_+^{A_n \times A_n}$. Associate the canonical component-wise order with X, with an exception that the order is reversed for $\omega \in \Omega^1$. X is a lattice, with a typical element (v, λ), where $v \in \mathbb{R}^\Omega$ and $\lambda \in \mathbb{R}_+^{A_1 \times A_1} \times \ldots \times \mathbb{R}_+^{A_n \times A_n}$.

The data-value problem is equivalent to the problem $\max_{(v,\lambda) \in S} f(v, \lambda; \mu)$, where $f(v, \lambda; \mu) = -\sum_{\omega \in \Omega} v(\omega)\mu(\omega)$ and the feasible set $S \subset X$ is given by the inequalities

$$v(\omega) \geq u_0(a, \omega) + \sum_{i \in I} \sum_{a_i' \in A_i} (u_i(a_i, a_{-i}, \omega) - u_i(a_i', a_{-i}, \omega))\lambda_i(a_i'|a_i).$$

We treat μ as a parameter. Note that S does not depend on μ. Furthermore, μ is an element of $(|\Omega| - 1)$-dimensional simplex, with which we associate the following partial order: $\mu' \geq \mu$ if $\mu'(\omega) \geq \mu(\omega)$ for $\omega \in \Omega^1$, $\mu'(\omega) \leq \mu(\omega)$ for $\omega \in \Omega^2$, and $\mu'(\omega) = \mu(\omega)$ for $\omega \in \Omega^3$. Note that $\mu_1 \geq \mu_2$ in accordance with this partial order.

We want to show that f is supermodular in (v, λ) and has increasing differences in $(v, \lambda; \mu)$. Observe that

$$\begin{aligned} f(v', \lambda'; \mu) + f(v'', \lambda''; \mu) &= -\sum_{\omega \in \Omega} v'(\omega)\mu(\omega) - \sum_{\omega \in \Omega} v''(\omega)\mu(\omega) \\ &= -\sum_{\omega \in \Omega} (v'(\omega) + v''(\omega))\mu(\omega) \\ &= -\sum_{\omega \in \Omega} (\max\{v'(\omega), v''(\omega)\} + \min\{v'(\omega), v''(\omega)\})\mu(\omega) \\ &= f((v', \lambda') \wedge (v'', \lambda''); \mu) + f((v', \lambda') \vee (v'', \lambda''); \mu). \end{aligned}$$

Then f is supermodular in (v, λ).

Fix $(v', \lambda') \geq (v, \lambda)$ and $\mu' \geq \mu$. Observe that

$$\begin{aligned} &(f(v', \lambda', \mu') - f(v, \lambda, \mu')) - (f(v', \lambda', \mu) - f(v, \lambda, \mu)) \\ = &\sum_{\omega \in \Omega} (v(\omega) - v'(\omega))(\mu'(\omega) - \mu(\omega)) \\ = &\sum_{\omega \in \Omega^1} (v(\omega) - v'(\omega))(\mu'(\omega) - \mu(\omega)) + \sum_{\omega \in \Omega^2} (v(\omega) - v'(\omega))(\mu'(\omega) - \mu(\omega)) \geq 0, \end{aligned}$$

where the inequality follows from the adapted partial orders. Then f has increasing differences in $(v, \lambda; \mu)$.

Finally, by Theorem 5 in Milgrom and Shannon (1994), $\arg\max_{(v,\lambda)\in S} f(v,\lambda;\mu)$ is monotone nondecreasing in μ. This monotone comparative statics coupled with generic uniqueness of (v_q^*, b_q^*, ℓ_q^*) with respect to q imply that if $\mu_q(\omega) > \mu_{q'}(\omega)$ for two databases q and q' then $v_q^*(\omega) \leq v_{q'}^*(\omega)$.

When only interactions of type ω are present in the database, that is, $\mu_q(\omega) = 1$, we have $v_q^*(\omega) = \bar{u}(\omega)$. Indeed, the definition of $\bar{u}(\omega)$ implies that it can be written as

$$\bar{u}(\omega) = \min_{b_\omega, \ell_\omega} \max_{a\in A} \{u_0(a,\omega) + t_{b_\omega,\ell_\omega}(a,\omega)\},$$

where $t_{b_\omega,\ell_\omega}(a,\omega) = \sum_{i\in I} b_{i,\omega}(a_i) \sum_{a_i'\in A_i}(u_i(a_i,a_{-i},\omega) - u_i(a_i',a_{-i},\omega))\ell_{i,\omega}(a_i'|a_i)$, $b_\omega = (b_{1,\omega},\ldots,b_{n,\omega})$, with $b_{i,\omega}: A_i \to \mathbb{R}_{++}$, and $\ell_\omega = (\ell_{1,\omega},\ldots,\ell_{n,\omega})$, with $\ell_{i,\omega}: A_i \to \Delta(A_i)$.

For $\varepsilon > 0$, consider a set $M_\varepsilon(\omega)$ defined as $M_\varepsilon(\omega) = \{\mu \in \Delta(\Omega) : \mu(\omega') \in (0,\varepsilon) \text{ for } \omega \neq \omega', \mu(\omega) < 1\}$. By Proposition 2, there exists a finite collection $\{\mathcal{P}_1,\ldots,\mathcal{P}_K\}$ of open, convex, and disjoint subsets of $\Delta(\Omega)$ such that $\cup_k \mathcal{P}_k$ has measure one and, for every k, (v_q^*, b_q^*, ℓ_q^*) is unique and constant for q, with $\mu_q \in \mathcal{P}_k$. Therefore, we can always find $\mathcal{P}_m \in \{\mathcal{P}_1,\ldots,\mathcal{P}_K\}$, such that $\mathcal{P}_m \cap M_\varepsilon(\omega)$ is nonempty, open, and convex for all $0 < \varepsilon \leq \delta$, where $\delta > 0$. Then $v_q^*(\omega)$ is unique and constant for all $q \in \mathbb{R}_{++}^\Omega$, with $\mu_q \in \mathcal{P}_m \cap M_\delta(\omega)$. Let us refer to this constant as $\hat{u}(\omega)$. If $\hat{u}(\omega) = \bar{u}(\omega)$, then the result follows. Suppose, on the contrary, that $\hat{u}(\omega) \neq \bar{u}(\omega)$. We can always pick a sequence μ^n, $n \in \mathbb{N}$, from $\mathcal{P}_m \cap M_\delta(\omega)$ that converges to $\tilde{\mu}$, with $\tilde{\mu}(\omega) = 1$. Then for every $n \in \mathbb{N}$, $v_q^*(\omega) = \hat{u}(\omega)$ for every q, such that $\mu_q = \mu^n$. By the Berge's maximum theorem, (v_q^*, b_q^*, ℓ_q^*) is an upper-hemicontinuous correspondence and therefore has closed graph. Hence, $\hat{u}(\omega) \in v_q^*(\omega)$ for every q, with $\mu_q = \tilde{\mu}$. We obtain the desired contradiction, since $v_q^*(\omega) = \bar{u}(\omega)$ for such q. □

Proof of Proposition 4. Fix $q \in \mathbb{R}_{++}^\Omega$. Suppose that a FIC mechanism x_q^* is optimal. Then, we have

$$v_q^*(\omega) = u_q^*(\omega) + \sum_{a\in A} t_q^*(a,\omega) x_q^*(a|\omega) \geq u_q^*(\omega),$$

where the inequality follows from $x_q^*(\cdot|\omega) \in CE(\Gamma_\omega)$ for all ω. Since by Lemma 1 we must have $\sum_\omega v_q^*(\omega) q(\omega) = \sum_\omega u_q^*(\omega) q(\omega)$, it follows that $v_q^*(\omega) = u_q^*(\omega)$ for all ω. Finally, since x_q^* is optimal, it must be that $u_q^*(\omega) = \bar{u}(\omega)$ for all ω. Now, note that v_q^* defines a supporting hyperplane of the iso-payoff line of level $U^*(q)$ at q. The intercept of such an hyperplane on each ω-axis is $\hat{q}_\omega(\omega) = \frac{U^*(q)}{\bar{u}(\omega)}$ and $\hat{q}_\omega(\omega') = 0$ for $\omega' \neq \omega$. By definition, each $\hat{q}_\omega$ also belongs to the iso-payoff line of level $U^*(q)$ and therefore $U^*(q) = U^*(\hat{q}_\omega)$ for all ω. In other words, the intercepts of the hyperplane and the iso-payoff line coincide for all ω.

Now consider any $q' \in \mathbb{R}_{++}$, $q' \neq q$, that belongs to the supporting hyperplane of level $U^*(q)$ at q. By definition, we can obtain q' as a convex combination of intercepts $\hat{q}_\omega$ on each axis.

Specifically, there exists $\beta \in \Delta(\Omega)$ such that $q'(\omega) = \beta(\omega)\hat{q}_\omega(\omega)$ for all ω. By concavity of $U^*(q)$ (Footnote 19), we must have that

$$U^*(q') = \sum_{\omega \in \Omega} v^*_{q'}(\omega) q'(\omega) \leq U^*(q) = \sum_{\omega \in \Omega} \beta(\omega) U^*(\hat{q}_\omega) = \sum_{\omega \in \Omega} \bar{u}(\omega) q'(\omega).$$

But since $v^*_{q'}(\omega) \geq \bar{u}(\omega)$ for all ω by Lemma 2, we must have $v^*_{q'}(\omega) = \bar{u}(\omega)$ for all ω. Then $v^*_{q''}(\omega) = \bar{u}(\omega)$ for all q'' that belong to the supporting hyperplane of level $U^*(q)$ at q. Finally, since v^*_q is invariant to scaling of q, it follows that $v^*_q(\omega) = \bar{u}(\omega)$ for all ω and all $q \in \mathbb{R}^\Omega_+$. □

Proof of Proposition 5. To build intuition, imagine the principal refines one ω-datapoint according to σ_ω, which does not change q by being infinitesimal. Since $u_i(a, \omega) = \mathbb{E}_\sigma[u_i(a, \omega')|\omega]$ for all $i = 0, 1, \ldots, n$, using expression (3), by complementary slackness we get

$$v^*_q(\omega) = \sum_{\omega' \in \Omega} [u_0(a^*_q(\omega), \omega') + t^*_q(a^*_q(\omega), \omega')]\sigma_\omega(\omega'),$$

where $a^*_q(\omega)$ is any action profile in the support of $x^*_q(\cdot|\omega)$. By (3) again, this implies that such a refinement increases the expected value of the refined ω-datapoint:

$$\sum_{\omega' \in \Omega} v^*_q(\omega')\sigma_\omega(\omega') - v^*_q(\omega) \geq 0. \tag{15}$$

Note that if refining $\alpha q(\omega)$ of the current ω-datapoints according to σ_ω does not change the value of datapoints, then (15) implies the desired inequality.

Now suppose that acquiring better data changes the value of datapoints. That is, there exists a share $\alpha > 0$ such that refining $\alpha q(\omega)$ of the current ω-datapoints according to σ_ω leads to a new database q_α such that $v^*_{q_\alpha}(\omega') \neq v^*_q(\omega')$ for some $\omega' \in \operatorname{supp} \sigma_\omega$ or $\omega' = \omega$. Since the total quantity of datapoints does not change, we have that $\mu_{q_\alpha}(\omega) < \mu_q(\omega)$ and $\mu_{q_\alpha}(\omega') > \mu_q(\omega')$ for all $\omega' \in \operatorname{supp} \sigma_\omega$. By Proposition 3, it follows that $v^*_{q_\alpha}(\omega) \geq v^*_q(\omega)$ and $v^*_{q_\alpha}(\omega') \leq v^*_q(\omega')$ for all $\omega' \in \operatorname{supp} \sigma_\omega$. Now, note that for all α,

$$\sum_{\omega' \in \Omega} v^*_{q_\alpha}(\omega')\sigma_\omega(\omega') \geq v^*_{q_\alpha}(\omega) \geq v^*_q(\omega), \tag{16}$$

where the first inequality follows from (15). This implies that the value of acquiring better data is always non-negative.

Now, suppose that there exists a common $\tilde{a} \in \operatorname{supp} x^*_q(\cdot|\omega)$ that satisfies $x^*_q(\tilde{a}|\omega'') > 0$ for all $\omega'' \in \operatorname{supp} \sigma_\omega$. By complementary slackness, it follows that for all $\omega'' \in \operatorname{supp} \sigma_\omega$, we have $v^*_q(\omega'') = u_0(\tilde{a}, \omega'') + t^*_q(\tilde{a}, \omega'')$. Therefore, by the scarcity principle,

$$\sum_{\omega'' \in \Omega} v^*_{q_\alpha}(\omega'')\sigma_\omega(\omega'') \leq \sum_{\omega'' \in \Omega} v^*_q(\omega'')\sigma_\omega(\omega'') = v^*_q(\omega) \leq v^*_{q_\alpha}(\omega),$$

which, combined with (16), implies the desired equality.

Conversely, suppose that for every $\hat{a} \in \operatorname{supp} x_q^*(\cdot|\omega)$ there exists $\omega' \in \operatorname{supp} \sigma_\omega$ that satisfies $x_q^*(\hat{a}|\omega') = 0$. If the solution to the data-value problem is unique for database q, then $x_q^*(\hat{a}|\omega') = 0$ implies $v_q^*(\omega') > u_0(\hat{a}, \omega') + t_q^*(\hat{a}, \omega')$ by strict complementary slackness. The desired strict inequality is then obtained.

Proof of Proposition 6. The directional derivative of U^* at q along the linear path from q to q_α is equal to

$$q(\omega)\left[\sum_{\omega' \in \Omega} v_q^*(\omega')\sigma_\omega(\omega') - v_q^*(\omega)\right].$$

The linear path from q to q_α can be parametrized as follows: for $t \in [0,1]$, define $q_t(\omega) = q(\omega) - t\alpha q(\omega)$, $q_t(\omega') = q(\omega') + t\alpha\sigma_\omega(\omega')q(\omega)$ for $\omega' \in \operatorname{supp} \sigma_\omega$, and $q_t(\omega'') = q(\omega'')$ for remaining ω''.

Note that $\sum_{\omega' \in \Omega} v_{q_t}^*(\omega')\sigma_\omega(\omega') - v_{q_t}^*(\omega)$ is non-negative by (15) and decreasing in t by the scarcity principle.

Finally, by the gradient theorem,

$$U^*(q_\alpha) - U^*(q) = \int_0^1 v_{q_t}^* \cdot \nabla q_t \, dt = \alpha q(\omega) \int_0^1 \left[\sum_{\omega' \in \Omega} v_{q_t}^*(\omega')\sigma_\omega(\omega') - v_{q_t}^*(\omega)\right] dt \geq 0,$$

where ∇q_t is the gradient of q_t with respect to t. □

Proof of Proposition 7 . We provide a proof for the general case where the principal can choose $a_0 \in A_0$ and each agent i can privately observe some own data $\omega_i \in \Omega_i$ about the interaction he is in. Fix $(a^*, \omega^*) \in \mathbf{X}$ and introduce $\mathbf{1}_{a^*,\omega^*}$ as a vector of size $|\mathbf{X}|$ with $\varepsilon > 0$ in the position indexed by (a^*, ω^*) and 0 in all other positions. Constitute a matrix $\mathbf{W}$ such that its rows are indexed by $(a, \omega) \in \mathbf{X}$, its columns are indexed by (i, a_i', a_i, ω_i), $i \in I$, and its entries are as follows:

$$\mathbf{W}\left((\tilde{a}, \tilde{\omega}), (i, a_i', a_i, \omega_i)\right) = 1\left\{a_i = \tilde{a}_i, \omega_i = \tilde{\omega}_i\right\}\left(u_i(a_i, a_{-i}, \omega_i) - u_i(a_i', a_{-i}, \omega_i)\right).$$

By a variant of the Farkas' lemma, either there exists $\lambda \geq 0$, such that $\mathbf{W}\lambda \leq -\mathbf{1}_{a^*,\omega^*}$, or else there exists $\chi \geq 0$, such that $\mathbf{W}^T\chi \geq 0$, with $\chi^T\mathbf{1}_{a^*,\omega^*} > 0$. Now we show that the latter is true. Indeed, we can pick $\chi(a, \omega) = q(\omega)x(a|\omega)$, where x is obedient and satisfies $x(a^*|\omega^*) > 0$. We can find such x, since $(a^*, \omega^*) \in \mathbf{X}$. Then $\chi \geq 0$ and $\chi^T\mathbf{1}_{a^*,\omega^*} > 0$ are satisfied automatically. Finally, $\mathbf{W}^T\chi \geq 0$ corresponds exactly to the set of obedience constraints in $\mathcal{U}_q$ restricted to the subdomain $\mathbf{X}$.

Since any λ can be decomposed as $\lambda_i(a_i'|a_i,\omega_i) = b_i(a_i,\omega_i)\ell_i(a_i'|a_i,\omega_i)$, we conclude that there is no $(b,\ell) \in G(\mathbf{X})$ that satisfies $t(a,\omega) \leq 0$ for every $(a,\omega) \in \mathbf{X}$ and $t(a^*,\omega^*) < -\varepsilon$. The result then follows, since the choice of $(a^*,\omega^*) \in \mathbf{X}$ and $\varepsilon > 0$ was arbitrary. □

Proof of Proposition 8. We will argue by contradiction. Suppose $q \in \mathbb{R}^{\Omega}_{++}$ and $\mathcal{U}_q$ admits an FIC solution x_q^{**} and hence $x_q^{**}(\cdot|\tilde{\omega}) \in CE(\Gamma_{\tilde{\omega}})$ and $u_q^{**}(\tilde{\omega}) = \bar{u}(\tilde{\omega})$ for all $\tilde{\omega} \in \Omega$. Then $v_q^{**}(\tilde{\omega}) = u_q^{**}(\tilde{\omega}) = \bar{u}(\tilde{\omega})$ for all $\tilde{\omega} \in \Omega$ by Proposition 4.

Now suppose that (a,ω) satisfies both conditions in the statement of the proposition. For $(v_q^{**}, b_q^{**}, \ell_q^{**})$ to be feasible for $\mathcal{V}_q$, we must have for all $\tilde{\omega} \in \Omega$,

$$v_q^{**}(\tilde{\omega}) \geq u_0(a,\tilde{\omega}) + t_q^{**}(a,\tilde{\omega}).$$

Since $u_0(a,\omega) > \bar{u}(\omega) = v^{**}(\omega)$, we must have $t_q^{**}(a,\omega) < 0$. Therefore, there exists a pair $(i,\hat{a}_i)$ that satisfies $u_i(a_i,a_{-i},\omega) < u_i(\hat{a}_i,a_{-i},\omega)$ and $\ell_{q,i}^{**}(\hat{a}_i|a_i,\omega_i) > 0$. For such a pair $(i,\hat{a}_i)$, there exists $x(\cdot|\omega') \in CE(\Gamma_{\omega'})$ with the properties listed in the proposition. Then, since $b_q^{**} > 0$ and $\ell_{q,i}^{**}(\hat{a}_i|a_i,\omega_i) > 0$,

$$\begin{aligned}
& \sum_{\tilde{a}\in A} u_0(\tilde{a},\omega')x(\tilde{a}|\omega') + \sum_{\tilde{a}\in A} t_q^{**}(\tilde{a},\omega')x(\tilde{a}|\omega') \\
\geq \; & \sum_{\tilde{a}\in A} u_0(\tilde{a},\omega')x(\tilde{a}|\omega') \\
& + b_{q,i}^{**}(a_i,\omega_i)\ell_{q,i}^{**}(\hat{a}_i|a_i,\omega_i)\left\{ \sum_{\tilde{a}_{-i}\in A_{-i}} \left(u_i(a_i,\tilde{a}_{-i},\omega') - u_i(\hat{a}_i,\tilde{a}_{-i},\omega')\right) x(a_i,\tilde{a}_{-i}|\omega') \right\} \\
> \; & \sum_{\tilde{a}\in A} u_0(\tilde{a},\omega')x(\tilde{a}|\omega') = v_q^{**}(\omega'),
\end{aligned}$$

where the first inequality follows because $x(\cdot|\omega') \in CE(\Gamma_{\omega'})$. The strict inequality is incompatible with constraint (3) and delivers the desired contradiction. □

C Analysis of the Leading Example

This section presents the calculations that back up our statements regarding the leading example. We can ignore the buyers and build their decisions into the utility functions of the surplus-maximizing platform ($i = 0$) and the seller ($i = 1$). There are three types of datapoints, labeled by $\omega \in \{\omega_L, \omega_H, \omega^\circ\}$, where $\omega_H > \omega_L > 0$, and corresponding to whether the buyer's revealed valuation is ω_L, ω_H, or unknown to the platform. Suppose ω° turns into ω_H with probability h and ω_L with probability $1-h$. The prices the seller can charge are $a \in \{\omega_L, \omega_H\}$. The payoffs are $u_0(a,\omega) = \max\{\omega - a, 0\}$, $u_1(a,\omega) = a$ if $a \leq \omega$, and $u_1(a,\omega) = 0$ if

$a > \omega$. Given this, we have $u_i(a, \omega^\circ) = hu_i(a, \omega_H) + (1-h)u_i(a, \omega_L)$ for $i = 0, 1$. For completeness, we solve both the information-design problem and the data-value problem separately. Since our goal here is to only find the optimizers in both problems, we can work in the space of databases that satisfy $q(\omega_L) + q(\omega_H) + q(\omega^\circ) = 1$, so that $q(\omega) = \mu_q(\omega)$.

C.1 Information-Design Problem

The objective of the platform is

$$(\omega_H - \omega_L)x(\omega_L|\omega_H)\mu_q(\omega_H) + h(\omega_H - \omega_L)x(\omega_L|\omega^\circ)\mu_q(\omega^\circ).$$

The obedience constraints are

$$-\omega_L x(\omega_H|\omega_L)\mu_q(\omega_L) + (\omega_H - \omega_L)x(\omega_H|\omega_H)\mu_q(\omega_H) + (h\omega_H - \omega_L)x(\omega_H|\omega^\circ)\mu_q(\omega^\circ) \geq 0,$$
$$\omega_L x(\omega_L|\omega_L)\mu_q(\omega_L) - (\omega_H - \omega_L)x(\omega_L|\omega_H)\mu_q(\omega_H) - (h\omega_H - \omega_L)x(\omega_L|\omega^\circ)\mu_q(\omega^\circ) \geq 0.$$

We consider two cases depending on whether $h\omega_H - \omega_L > 0$, or $h\omega_H - \omega_L \leq 0$. If $h\omega_H - \omega_L > 0$, or $h > \frac{\omega_L}{\omega_H}$, then from the second obedience constraint $x_q^*(\omega_L|\omega_L) = 1$. The first obedience constraint is then automatically satisfied. Since $h \in (0, 1)$, it is always true that $\frac{h\omega_H - \omega_L}{\omega_H - \omega_L} < h$. Then the solution satisfies $x_q^*(\omega_L|\omega_H) = 0$ and $x_q^*(\omega_L|\omega^\circ) = \frac{\omega_L}{h\omega_H - \omega_L}\frac{\mu_q(\omega_L)}{\mu_q(\omega^\circ)}$, as long as $\frac{\omega_L}{h\omega_H - \omega_L}\frac{\mu_q(\omega_L)}{\mu_q(\omega^\circ)} \leq 1$. We conclude that the solution is as follows:

1. If $\mu_q(\omega_L) \leq \frac{h\omega_H - \omega_L}{\omega_L}\mu_q(\omega^\circ)$, then
$$x_q^*(\omega_L|\omega_L) = 1, x_q^*(\omega_L|\omega_H) = 0, \text{ and } x_q^*(\omega_L|\omega^\circ) = \frac{\omega_L}{h\omega_H - \omega_L}\frac{\mu_q(\omega_L)}{\mu_q(\omega^\circ)};$$
2. If $\frac{\omega_H - \omega_L}{\omega_H} - \mu_q(\omega^\circ)(1-h) \geq \mu_q(\omega_L) \geq \frac{h\omega_H - \omega_L}{\omega_L}\mu_q(\omega^\circ)$, then
$$x_q^*(\omega_L|\omega_L) = 1, x_q^*(\omega_L|\omega_H) = \frac{\omega_L\mu_q(\omega_L) - (h\omega_H - \omega_L)\mu_q(\omega^\circ)}{(\omega_H - \omega_L)\mu_q(\omega_H)}, \text{ and } x_q^*(\omega_L|\omega^\circ) = 1;$$
3. If $\mu_q(\omega_L) \geq \frac{\omega_H - \omega_L}{\omega_H} - \mu_q(\omega^\circ)(1-h)$, then
$$x_q^*(\omega_L|\omega_L) = 1, x_q^*(\omega_L|\omega_H) = 1, \text{ and } x_q^*(\omega_L|\omega^\circ) = 1.$$

Now suppose that $h\omega_H - \omega_L \leq 0$, or $h \leq \frac{\omega_L}{\omega_H}$. Combining obedience constraints in the standard manner for communication problems with binary action, we get

$$\omega_L x(\omega_L|\omega_L)\mu_q(\omega_L) - (\omega_H - \omega_L)x(\omega_L|\omega_H)\mu_q(\omega_H) - (h\omega_H - \omega_L)x(\omega_L|\omega^\circ)\mu_q(\omega^\circ) \geq$$

$$\max\left\{\omega_H\mu_q(\omega_L) + (1-h)\omega_H\mu_q(\omega^\circ) - (\omega_H - \omega_L), 0\right\}.$$

It is immediate that $x_q^*(\omega_L|\omega^\circ) = x_q^*(\omega_L|\omega_L) = 1$, since this choice relaxes the platform's problem as much as possible. The obedience constraint then becomes

$$\omega_L\mu_q(\omega_L) - (h\omega_H - \omega_L)\mu_q(\omega^\circ) - \max\left\{\omega_H\mu_q(\omega_L) + (1-h)\omega_H\mu_q(\omega^\circ) - (\omega_H - \omega_L), 0\right\} \geq$$

$$(\omega_H - \omega_L)x(\omega_L|\omega_H)\mu_q(\omega_H)$$

We then conclude that the solution is as follows:

1. If $\mu_q(\omega_L) \leq \frac{\omega_H - \omega_L}{\omega_H} - \mu_q(\omega^\circ)(1-h)$, then

$$x_q^*(\omega_L|\omega_L) = 1, x_q^*(\omega_L|\omega_H) = \frac{\omega_L\mu_q(\omega_L) - (h\omega_H - \omega_L)\mu_q(\omega^\circ)}{(\omega_H - \omega_L)\mu_q(\omega_H)}, \text{ and } x_q^*(\omega_L|\omega^\circ) = 1;$$

2. If $\mu_q(\omega_L) \geq \frac{\omega_H - \omega_L}{\omega_H} - \mu_q(\omega^\circ)(1-h)$, then

$$x_q^*(\omega_L|\omega_L) = 1, x_q^*(\omega_L|\omega_H) = 1, \text{ and } x_q^*(\omega_L|\omega^\circ) = 1.$$

C.2 Data-Value Problem

Let $\lambda(a_1'|a_1) = b(a_1)l(a_1'|a_1)$. The data-value problem is then

$$\min_{v,\lambda} \mu_q(\omega_L)v(\omega_L) + \mu_q(\omega_H)v(\omega_H) + \mu_q(\omega^\circ)v(\omega^\circ),$$

subject to $\lambda(\omega_L|\omega_H), \lambda(\omega_H|\omega_L) \geq 0$,

$$\begin{aligned}
v(\omega_L) &= \max\{\omega_L\lambda(\omega_L|\omega_H), -\omega_L\lambda(\omega_H|\omega_L)\} = \omega_L\lambda(\omega_L|\omega_H),\\
v(\omega_H) &= \max\{\omega_H - \omega_L - (\omega_H - \omega_L)\lambda(\omega_L|\omega_H), (\omega_H - \omega_L)\lambda(\omega_H|\omega_L)\} =\\
&\quad (\omega_H - \omega_L)\max\{1 - \lambda(\omega_L|\omega_H), \lambda(\omega_H|\omega_L)\},\\
v(\omega^\circ) &= \max\{h(\omega_H - \omega_L) + (\omega_L - h\omega_H)\lambda(\omega_L|\omega_H), (h\omega_H - \omega_L)\lambda(\omega_H|\omega_L)\} =\\
&\quad h(\omega_H - \omega_L)\max\left\{1 - \frac{h\omega_H - \omega_L}{h(\omega_H - \omega_L)}\lambda(\omega_L|\omega_H), \frac{h\omega_H - \omega_L}{h(\omega_H - \omega_L)}\lambda(\omega_H|\omega_L)\right\}.
\end{aligned}$$

As we noted before, $\frac{h\omega_H - \omega_L}{h(\omega_H - \omega_L)} < 1$. Suppose that $h > \frac{\omega_L}{\omega_H}$. Then it is optimal to set $\lambda_q^*(\omega_H|\omega_L) = 0$ to relax the problem as much as possible. We then have

$$\begin{aligned}
v(\omega_L) &= \omega_L\lambda(\omega_L|\omega_H),\\
v(\omega_H) &= (\omega_H - \omega_L)\max\{1 - \lambda(\omega_L|\omega_H), 0\},
\end{aligned}$$

$$v(\omega^\circ) = h(\omega_H - \omega_L) \max\left\{1 - \frac{h\omega_H - \omega_L}{h(\omega_H - \omega_L)}\lambda(\omega_L|\omega_H), 0\right\}.$$

There are three candidates for optimal $\lambda(\omega_L|\omega_H)$, specifically, 0 and two kinks of the maxima in the expressions above, 1 and $\frac{h(\omega_H-\omega_L)}{h\omega_H-\omega_L} > 1$.

When $\lambda(\omega_L|\omega_H) = 0$, the objective is $S_0 := (1 - \mu_q(\omega_L) - \mu_q(\omega^\circ)(1-h))(\omega_H - \omega_L)$.

When $\lambda(\omega_L|\omega_H) = 1$, the objective is $S_1 := \mu_q(\omega_L)\omega_L + \mu_q(\omega^\circ)(1-h)\omega_L$.

When $\lambda(\omega_L|\omega_H) = \frac{h(\omega_H-\omega_L)}{h\omega_H-\omega_L}$, the objective is $S_f := \mu_q(\omega_L)\frac{h(\omega_H-\omega_L)}{h\omega_H-\omega_L}\omega_L$.

The following claims are true:

- $S_0 \leq S_1$ if and only if $\mu_q(\omega_L) + \mu_q(\omega^\circ)(1-h) \geq \frac{\omega_H-\omega_L}{\omega_H}$;
- $S_0 \leq S_f$ if and only if $\mu_q(\omega_L)\left(1 + \frac{h\omega_L}{h\omega_H-\omega_L}\right) + \mu_q(\omega^\circ)(1-h) \geq 1$;
- $S_1 \leq S_f$ if and only if $\mu_q(\omega_L)\frac{\omega_L}{h\omega_H-\omega_L} \geq \mu_q(\omega^\circ)$.

Figure 2 captures the resulting regions of $\mu_q(\omega_L)$ and $\mu_q(\omega^\circ)$ that correspond to the value of the problem being equal to one of S_0, S_1, and S_f.

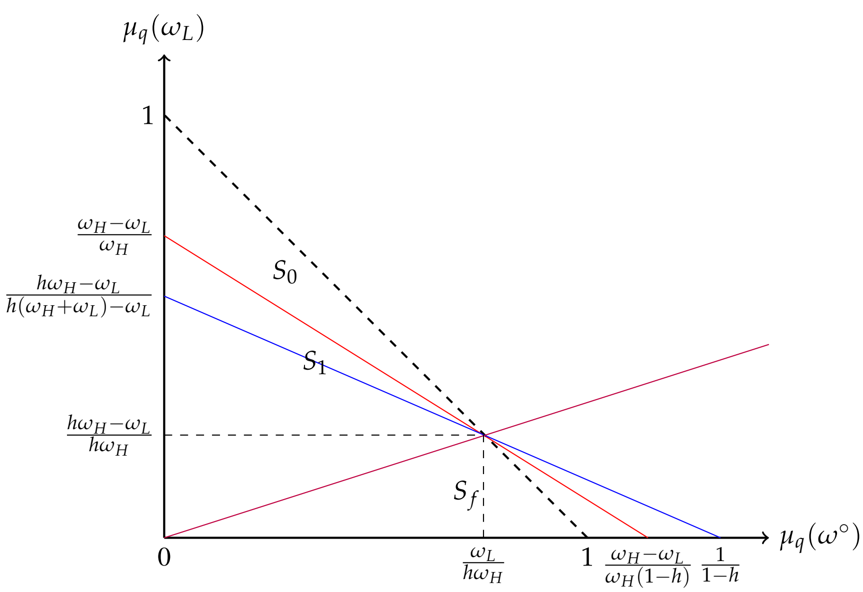


Figure 2: This figure pins down the minimum value of the data-value problem depending on μ_q when $h > \frac{\omega_L}{\omega_H}$. The red line corresponds to $S_0 = S_1$, the blue line corresponds to $S_0 = S_f$, and the purple line corresponds to $S_1 = S_f$.

We then conclude that the solution is as follows:

1. If $\mu_q(\omega_L) \geq \frac{\omega_H - \omega_L}{\omega_H} - \mu_q(\omega^\circ)(1-h)$, then $\lambda_q^*(\omega_L|\omega_H) = 0$ is optimal. The resulting values are

$$v_q^*(\omega_L) = 0, v_q^*(\omega_H) = \omega_H - \omega_L, \text{ and } v_q^*(\omega^\circ) = h(\omega_H - \omega_L);$$

2. If $\frac{h\omega_H - \omega_L}{\omega_L}\mu_q(\omega^\circ) \leq \mu_q(\omega_L) \leq \frac{\omega_H - \omega_L}{\omega_H} - \mu_q(\omega^\circ)(1-h)$, then $\lambda_q^*(\omega_L|\omega_H) = 1$ is optimal. The resulting values are

$$v_q^*(\omega_L) = \omega_L, v_q^*(\omega_H) = 0, \text{ and } v_q^*(\omega^\circ) = (1-h)\omega_L;$$

3. If $\mu_q(\omega_L) \leq \frac{h\omega_H - \omega_L}{\omega_L}\mu_q(\omega^\circ)$, then $\lambda_q^*(\omega_L|\omega_H) = \frac{h(\omega_H - \omega_L)}{h\omega_H - \omega_L}$ is optimal. The resulting values are

$$v_q^*(\omega_L) = \frac{h(\omega_H - \omega_L)}{h\omega_H - \omega_L}\omega_L, v_q^*(\omega_H) = 0, \text{ and } v_q^*(\omega^\circ) = 0.$$

Suppose now that $h \leq \frac{\omega_L}{\omega_H}$. Then immediately

$$v(\omega^\circ) = h(\omega_H - \omega_L) - (h\omega_H - \omega_L)\lambda(\omega_L|\omega_H).$$

$\lambda_q^*(\omega_H|\omega_L) = 0$ is optimal again. There are only two candidates for optimal $\lambda(\omega_L|\omega_H)$, specifically, 0 and 1. The solution then is as follows:

1. If $\mu_q(\omega_L) \geq \frac{\omega_H - \omega_L}{\omega_H} - \mu_q(\omega^\circ)(1-h)$, then $\lambda_q^*(\omega_L|\omega_H) = 0$ is optimal. The resulting values are

$$v_q^*(\omega_L) = 0, v_q^*(\omega_H) = \omega_H - \omega_L, \text{ and } v_q^*(\omega^\circ) = h(\omega_H - \omega_L);$$

2. If $\mu_q(\omega_L) \leq \frac{\omega_H - \omega_L}{\omega_H} - \mu_q(\omega^\circ)(1-h)$, then $\lambda_q^*(\omega_L|\omega_H) = 1$ is optimal. The resulting values are

$$v_q^*(\omega_L) = \omega_L, v_q^*(\omega_H) = 0, \text{ and } v_q^*(\omega^\circ) = (1-h)\omega_L.$$

C.3 Summary

All the cases considered can be grouped into three scenarios based on $\mu_q(\omega_L)$ and $\mu_q(\omega^\circ)$.

Scenario 1. Suppose that $\mu_q(\omega_L) \leq \frac{h\omega_H - \omega_L}{\omega_L}\mu_q(\omega^\circ)$. This scenario corresponds to condition (i) of Table 1. Note that this scenario appears only if $h > \frac{\omega_L}{\omega_H}$. The solution to the information-design problem is presented in Table 3.

$x_q^*(a\|\omega)$		ω		
		ω_L	ω_H	ω°
a	ω_L	1	0	$\frac{\omega_L}{h\omega_H-\omega_L}\frac{\mu_q(\omega_L)}{\mu_q(\omega^\circ)}$
	ω_H	0	1	$1-\frac{\omega_L}{h\omega_H-\omega_L}\frac{\mu_q(\omega_L)}{\mu_q(\omega^\circ)}$

Table 3: Platform Example, x_q^* for Scenario 1.

The solution to the data-value problem is $\lambda_q^*(\omega_H|\omega_L) = 0$, $\lambda_q^*(\omega_L|\omega_H) = \frac{h(\omega_H-\omega_L)}{h\omega_H-\omega_L}$ and the unit values of datapoints are $v_q^*(\omega_L) = \frac{h(\omega_H-\omega_L)}{h\omega_H-\omega_L}\omega_L$, $v_q^*(\omega_H) = 0$, and $v_q^*(\omega^\circ) = 0$.

Scenario 2. Suppose that $\frac{h\omega_H-\omega_L}{\omega_L}\mu_q(\omega^\circ) \leq \mu_q(\omega_L) \leq \frac{\omega_H-\omega_L}{\omega_H} - \mu_q(\omega^\circ)(1-h)$. This scenario corresponds to condition (ii) of Table 1. Note that the lower bound on $\mu_q(\omega_L)$ is meaningful only if $h > \frac{\omega_L}{\omega_H}$. The solution to the information-design problem is presented in Table 4.

$x_q^*(a\|\omega)$		ω		
		ω_L	ω_H	ω°
a	ω_L	1	$\frac{\omega_L\mu_q(\omega_L)-(h\omega_H-\omega_L)\mu_q(\omega^\circ)}{(\omega_H-\omega_L)\mu_q(\omega_H)}$	1
	ω_H	0	$\frac{\omega_H-\omega_L-\mu_q(\omega_L)\omega_H-\mu_q(\omega^\circ)(1-h)\omega_H}{(\omega_H-\omega_L)\mu_q(\omega_H)}$	0

Table 4: Platform Example, x_q^* for Scenario 2.

The solution to the data-value problem is $\lambda_q^*(\omega_H|\omega_L) = 0$, $\lambda_q^*(\omega_L|\omega_H) = 1$, and the unit values of datapoints are $v_q^*(\omega_L) = \omega_L$, $v_q^*(\omega_H) = 0$, and $v_q^*(\omega^\circ) = (1-h)\omega_L$.

Scenario 3. Suppose that $\mu_q(\omega_L) \geq \frac{\omega_H-\omega_L}{\omega_H} - \mu_q(\omega^\circ)(1-h)$. This scenario corresponds to condition (iii) of Table 1. The solution to the information-design problem is presented in Table 5.

$x_q^*(a\|\omega)$		ω		
		ω_L	ω_H	ω°
a	ω_L	1	1	1
	ω_H	0	0	0

Table 5: Platform Example, x_q^* for Scenario 3.

The solution to the data-value problem is $\lambda_q^*(\omega_H|\omega_L) = \lambda_q^*(\omega_L|\omega_H) = 0$ and the unit values of datapoints are $v_q^*(\omega_L) = 0$, $v_q^*(\omega_H) = \omega_H - \omega_L$, and $v_q^*(\omega^\circ) = h(\omega_H - \omega_L)$. $\triangle$